Analyzing the tree-layer structure of Deep Forests

Ludovic Arnould, Claire Boyer, Erwan Scornet

Analyzing the tree-layer structure of Deep Forests

Ludovic Arnould[1], Claire Boyer[1], and Erwan Scornet[2]

[1]LPSM, Sorbonne Université
[2]CMAP, Ecole Polytechnique

March 18, 2021

Abstract

Random forests on the one hand, and neural networks on the other hand, have met great success in the machine learning community for their predictive performance. Combinations of both have been proposed in the literature, notably leading to the so-called deep forests (DF) [28]. In this paper, our aim is not to benchmark DF performances but to investigate instead their underlying mechanisms. Additionally, DF architecture can be generally simplified into more simple and computationally efficient shallow forests networks. Despite some instability, the latter may outperform standard predictive tree-based methods. We exhibit a theoretical framework in which a shallow tree network is shown to enhance the performance of classical decision trees. In such a setting, we provide tight theoretical lower and upper bounds on its excess risk. These theoretical results show the interest of tree-network architectures for well-structured data provided that the first layer, acting as a data encoder, is rich enough.

1 Introduction

Deep Neural Networks (DNNs) are among the most widely used machine learning algorithms. They are composed of parameterized differentiable non-linear modules trained by gradient-based methods, which rely on the backpropagation procedure. Their performance mainly relies on layer-by-layer processing as well as feature transformation across layers. Training neural networks usually requires complex hyper-parameter tuning [1] and a huge amount of data. Although DNNs recently achieved great results in many areas, they remain very complex to handle, unstable to input noise [27] and difficult to interpret [17].

Recently, several attempts have been made to consider networks with non-differentiable modules. Among them the Deep Forest (DF) algorithm [28], which uses Random Forests (RF) [6] as neurons, has received a lot of attention in recent years in various applications such as hyperspectral image processing [16], medical imaging [22], drug interactions [21, 25] or even fraud detection [26].

Since the DF procedure stacks multiple layers, each one being composed of complex nonparametric RF estimators, the rationale behind the procedure remains quite obscure. However DF methods exhibit impressive performances in practice, suggesting that stacking RFs and extracting features from these estimators at each layer is a promising way to leverage on the RF performance in the neural network framework. The goal of this paper is not an exhaustive empirical study of prediction performances of DF [see 29] but rather to understand how stacking trees in a network fashion may result in competitive infrastructure.

Related Works. Different manners of stacking trees exist, as the Forwarding Thinking Deep Random Forest (FTDRF), proposed by [18], for which the proposed network contains trees which directly transmit their output to the next layer (contrary to deep forest in which their output is first averaged before being passed to the next layer). A different approach by [9] consists in rewriting tree gradient boosting as a simple neural network whose layers can be made arbitrary large depending on the boosting tree structure. The resulting estimator is more simple than DF but does not leverage on the ensemble method properties of random forests.

In order to prevent overfitting and to lighten the model, several ways to simplify DF architecture have been investigated. [19] considers RF whose complexity varies through the network, and combines

it with a confidence measure to pass high confidence instances directly to the output layer. Other directions towards DF architecture simplification are to play on the nature of the RF involved [3] (using Extra-Trees instead of Breiman's RF), on the number of RF per layer [13] (implementing layers of many forests with few trees), or even on the number of features passed between two consecutive layers [21] by relying on an importance measure to process only the most important features at each level. The simplification can also occur once the DF architecture is trained, as in [14] selecting in each forest the most important paths to reduce the network time- and memory-complexity. Approaches to increase the approximation capacity of DF have also been proposed by adjoining weights to trees or to forests in each layer [23, 24], replacing the forest by more complex estimators (cascade of ExtraTrees) [2], or by combining several of the previous modifications notably incorporating data preprocessing [12]. Overall, the related works on DF exclusively represent algorithmic contributions without a formal understanding of the driving mechanisms at work inside the forest cascade.

Contributions. In this paper, we analyze the benefit of combining trees in network architecture both theoretically and numerically. As the performances of DF have already been validated by the literature [see 29], the main goals of our study are (i) to quantify the potential benefits of DF over RF, and (ii) to understand the mechanisms at work in such complex architectures. We show in particular that much lighter configuration can be on par with DF default configuration, leading to a drastic reduction of the number of parameters in few cases. For most datasets, considering DF with two layers is already an improvement over the basic RF algorithm. However, the performance of the overall method is highly dependent on the structure of the first random forests, which leads to stability issues. By establishing tight lower and upper bounds on the risk, we prove that a shallow tree-network may outperform an individual tree in the specific case of a well-structured dataset if the first encoding tree is rich enough. This is a first step to understand the interest of extracting features from trees, and more generally the benefit of tree networks.

Agenda. DF are formally described in Section 2. Section 3 is devoted to the numerical study of DF, by evaluating the influence of the number of layers in DF architecture, by showing that shallow sub-models of one or two layers perform the best, and finally by understanding the influence of tree depth in cascade of trees. Section 4 contains the theoretical analysis of the shallow centered tree network. For reproducibility purposes, all codes together with all experimental procedures are to be found in the supplementary materials.

2 Deep Forests

2.1 Description

Deep Forest [28] is a hybrid learning procedure in which random forests are used as the elementary components (neurons) of a neural network. Each layer of DF is composed of an assortment of Breiman's forests and Completely-Random Forests (CRF) [8] and trained one by one. In a classification setting, each forest of each layer outputs a class probability distribution for any query point x, corresponding to the distribution of the labels in the node containing x. At a given layer, the distributions output by all forests of this layer are concatenated, together with the raw data. This new vector serves as input for the next DF layer. This process is repeated for each layer and the final classification is performed by averaging the forest outputs of the best layer (without raw data) and applying the `argmax` function. The overall architecture is depicted in Figure 1.

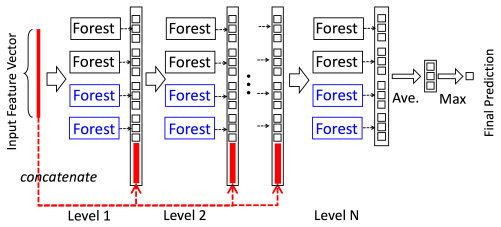


Figure 1: Deep Forest architecture (the scheme is taken from [28]).

2.2 DF hyperparameters

Deep Forests contain an important number of tuning parameters. Apart from the traditional parameters of random forests, DF architecture depends on the number of layers, the number of forests per layer, the type and proportion of random forests to use (Breiman or CRF). In [28], the default configuration is set to 8 forests per layer, 4 CRF and 4 RF, 500 trees per forest (other forest parameters are set to `sk-learn` [20] default values), and layers are added until 3 consecutive layers do not show score improvement.

Due to their large number of parameters and the fact that they use a complex algorithm as elementary bricks, DF consist in a potential high-capacity procedure. However, as a direct consequence, the numerous parameters are difficult to estimate (requiring specific tuning of the optimization process) and need to be stored which leads to high prediction time and large memory consumption. Besides, the layered structure of this estimate, and the fact that each neuron is replaced by a powerful learning algorithm makes the whole prediction hard to properly interpret.

As already pointed out, several attempts to lighten the architecture have been conducted. In this paper, we will propose and assess the performance of a lighter DF configuration on tabular datasets.

Remark 1. *DF [28] was first designed to classify images. To do so, a pre-processing network called Multi Grained Scanning (MGS) based on convolutions is first applied to the original images. Then the Deep Forest algorithm runs with the newly created features as inputs.*

3 Refined numerical analysis of DF architectures

In order to understand the benefit of using a complex architecture like Deep Forests, we compare different configurations of DF on six datasets in which the output is binary, multi-class or continuous, see Table 1 for description. All classification datasets belong to the UCI repository, the two regression ones are Kaggle datasets (Housing data and Airbnb Berlin 2020) [1].

Dataset	Type (Nb of classes)	Train/Test Size	Dim
Adult	Class. (2)	32560 / 16281	14
Higgs	Class. (2)	120000 / 80000	28
Letter	Class. (26)	16000 / 4000	16
Yeast	Class. (10)	1038 / 446	8
Airbnb	Regr.	91306 / 39132	13
Housing	Regr.	1095/ 365	61

Table 1: Description of the datasets.

In what follows, we propose a light DF configuration. We show that our light configuration performance is comparable to the performance of the default DF architecture of [28], thus questioning the relevance of deep models. Therefore, we analyze the influence of the number of layers in DF architectures, showing that DF improvements mostly rely on the first layers of the architecture. To gain insights about the quality of the new features created by the first layer, we consider a shallow tree network for which we evaluate the performance as a function of the first-tree depth.

3.1 Towards DF simplification

Setting. We compare the performances of the following DF architectures on the datasets summarized in Table 1:

(i) the default setting of DF, described in Section 2,

(ii) the best DF architecture obtained by grid-searching over the number of forests per layer, the number of trees per forest, and the maximum depth of each tree;

(iii) a new light DF architecture, composed of 2 layers, 2 forests per layer (one RF and one CRF) with only 50 trees of depth 30 trained only once.

[1] https://www.kaggle.com/raghavs1003/airbnb-berlin-2020
https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data

Results. Results are presented in Figures 2 and 3. Each bar plot respectively corresponds to the average accuracy or the average R^2 score over 10 tries for each test dataset; the error bars stand for accuracy or R^2 standard deviation. The description of the resulting best DF architecture for each dataset is given in Table S3 (Supplementary Materials). As highlighted in Figure 2, the

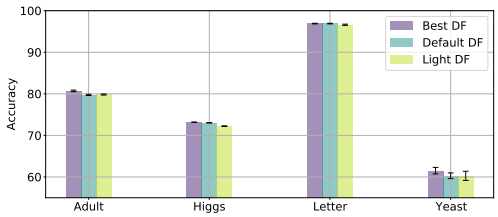


Figure 2: Accuracy of different DF architectures for classification datasets (10 runs per bar plot).

Figure 3: R^2 score of different DF architectures for regression datasets (10 runs per bar plot).

performance of the light configuration for classification datasets is comparable to the default and the best configurations, while being much more computationally efficient: faster to train, faster at prediction, cheaper in terms of memory (see Table S2 in the Supplementary Materials for details). This should be qualified by the yardstick of dataset regression results (see Figure 3). Indeed, for this type of problems, each forest in each layer outputs a scalar compared to the classification tasks in which the output is a vector whose size equals the number of classes. Therefore in regression, the extracted representation at each layer is simplistic thus requiring a deeper architecture.

Overall, for classification tasks, the small performance enhancement of deep forests (Default or Best DF) over our light configuration should be assessed in the light of their additional complexity. This questions the usefulness of stacking several layers made of many forests, resulting into a heavy architecture. We further propose an in-depth analysis of the role of each layer to the global DF performance.

3.2 A precise understanding of depth enhancement

In order to finely grasp the influence of tree depth in DF, we study a simplified version: a shallow CART tree network, composed of two layers, with one CART per layer.

Setting. In such an architecture, the first-layer tree is fitted on the training data. For each sample, the first-layer tree outputs a probability distribution (or a value in a regression setting), which is referred to as "encoded data" and given as input to the second-layer tree, with the raw features as well. For instance, considering binary classification data with classes 0 and 1, with raw features (x_1, x_2, x_3), the input of the second-layer tree is a 5-dimensional feature vector $(x_1, x_2, x_3, p_0, p_1)$, with p_0 (resp. p_1) the predicted probabilities by the first-layer tree for the class 0 (resp. 1).

For each dataset of Table 1, we first determine the optimal depth $k^\star$ of a single CART tree via 3-fold cross validation. Then, for a given first-layer tree with a fixed depth, we fit a second-layer tree, allowing its depth to vary. We then compare the resulting shallow tree networks in three different cases: when the (fixed) depth of the first tree is (i) less than $k^\star$, (ii) equal to $k^\star$, and (iii) larger than $k^\star$. We add the optimal single tree performance to the comparison.

Results. Results are displayed in Figure 4 for the Adult dataset only (see Supplementary Materials S1.3 for the results on the other datasets). Specifically noticeable in Figure 4 (top), the tree network architecture can introduce performance instability when the second-layer tree grows (e.g. when the latter is successively of depth 7, 8 and 9).

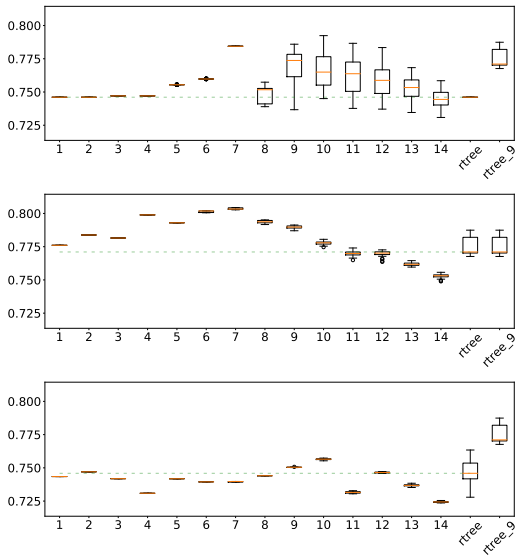


Figure 4: Adult dataset. Accuracy of a two-layer tree architecture w.r.t. the second-layer tree depth, when the first-layer (encoding) tree is of depth 2 (top), 9 (middle), and 15 (bottom). `rtree` is a single tree of respective depth 2 (top), 9 (middle), and 15 (bottom), applied on raw data. For this dataset, the optimal depth of a single tree is 9 and the tree with the optimal depth is depicted as `rtree_9` in each plot. The green dashed line indicates the median score of the `rtree`. All boxplots are obtained by 10 different runs.

Furthermore, when the encoding tree is not deep enough (top), the second-layer tree improves the accuracy until it approximately reaches the optimal depth $k^\star$. In this case, the second-layer tree compensates for the poor encoding, but cannot improve over a single tree with optimal depth $k^\star$. Conversely, when the encoding tree is more developed than an optimal single tree (bottom) - overfitting regime, the second-layer tree may not lead to any improvement, or worse, may degrade the performance of the first-layer tree.

On all datasets, the second-layer tree is observed to always make its first cut over the new features (see Figure 5 and Supplementary Materials).

In the case of binary classification, a single cut of the second-layer tree along a new feature yields to gather all the leaves of the first tree, predicted respectively as 0 and 1, into two big leaves, therefore reducing the predictor variance (cf. Figure 4 (middle and bottom)). Furthermore, when considering multi-label classification with n_{classes}, the second-layer tree must cut over at least n_{classes} features to recover the partition of the first tree (see Figure S13). Similarly, in the regression case, the second tree needs to perform a number of splits equal to the number of leaves of the first tree in order to recover the partition of the latter.

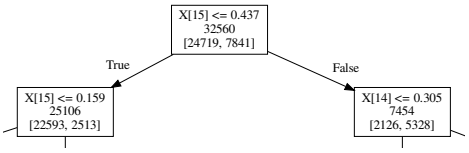


Figure 5: Adult dataset. Focus on the first levels of the second-layer tree structure when the first layer tree is of depth 9 (optimal depth). Raw features range from X[0] to X[13], X[14] and X[15] are the features built by the first-layer tree.

In Figure 4 (middle), one observes that with a first-layer tree of optimal depth, the second-layer tree may outperform an optimal single tree, by improving both the average accuracy and its variance. We aim at theoretically quantifying this performance gain in the next section.

4 Theoretical study of a shallow tree network

In this section, we focus on the theoretical analysis of a simplified tree network. Our aim is to exhibit settings in which a tree network outperforms a single tree. Recall that the second layer of a tree network gathers tree leaves of the first layer with similar distributions. For this reason, we believe that a tree network is to be used when the dataset has a very specific structure, in which the same link between the input and the output can be observed in different subareas of the input space. Such a setting is described in Section 4.2

To make the theoretical analysis possible, we study centered trees (see Definition 1) instead of CART. Indeed, studying the original CART algorithm is still nowadays a real challenge and analyzing stacks of CART seems out-of-reach in short term. As highlighted by the previous empirical analysis, we believe that the results we establish theoretically are shared by DF. All proofs are postponed to the Supplementary Materials.

4.1 The network architecture

We assume to have access to a dataset $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ of i.i.d. copies of the generic pair (X, Y).

Notations. Given a decision tree, we denote by $L_n(X)$ the leaf of the tree containing X and $N_n(L_n(X))$ the number of data points falling into $L_n(X)$. The prediction of such a tree at point X is given by

$$\hat{r}_n(X) = \frac{1}{N_n(L_n(X))} \sum_{X_i \in L_n(X)} Y_i$$

with the convention $0/0 = 0$, i.e. the prediction for X in a leaf with no observations is arbitrarily set to zero.

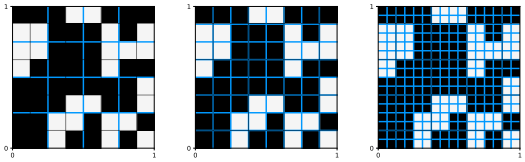
A shallow centered tree network. We want to theoretically analyze the benefits of stacking trees. To do so, we focus on two trees in cascade and will try to determine, in particular, the influence of the first (encoding) tree on the performance of the whole tree network. To catch the variance reduction property of tree networks already emphasized in the previous section, we consider a regression setting: let $r(x) = \mathbb{E}[Y|X = x]$ be the regression function and for any function f, its quadratic risk is defined as $R(f) = \mathbb{E}[(f(X) - r(X))^2]$, where the expectation is taken over $(X, Y, \mathcal{D}_n)$.

Definition 1 (Shallow centered tree network)**.** *The shallow tree network consists in two trees in cascade:*

- ***(Encoding layer)*** *The first-layer tree is a* cycling centered tree of depth k. *It is built independently of the data by splitting recursively on each variable, at the center of the cells. The tree construction is stopped when all cells have been cut exactly k times along each variable. For each point X, we extract the empirical mean $\bar{Y}_{L_n(X)}$ of the outputs Y_i falling into the leaf $L_n(X)$ and we pass the new feature $\bar{Y}_{L_n(X)}$ to the next layer, together with the original features X.*
- ***(Output layer)*** *The second-layer tree is a* centered tree of depth k' *for which a cut can be performed at the center of a cell along a raw feature (as done by the encoding tree) or along the new feature $\bar{Y}_{L_n(X)}$. In this latter case, two cells corresponding to $\{\bar{Y}_{L_n(X)} < 1/2\}$ and $\{\bar{Y}_{L_n(X)} \geq 1/2\}$ are created.*

The resulting predictor composed of the two trees in cascade, of respective depth k and k', trained on the data $(X_1, Y_1), \ldots, (X_n, Y_n)$ is denoted by $\hat{r}_{k,k',n}$.

The two cascading trees can be seen as two layers of trees, hence the name of the shallow tree network. Note in particular that $\hat{r}_{k,0,n}(X)$ is the prediction given by the first encoding tree only and outputs, as a classical tree, the mean of the Y_i's falling into a leaf containing X.



(a) Depth 4 (b) Depth 6 (c) Depth 8

Figure 6: Arbitrary chessboard data distribution for $k^\star = 6$ and $N_{\mathcal{B}} = 40$ black cells (p is not displayed here). Partition of the (first) encoding tree of depth 4, 6, 8 (from left to right) is displayed in blue. The optimal depth of a single centered tree for this chessboard distribution is 6.

4.2 Problem setting

Data generation. The data X is assumed to be uniformly distributed over $[0,1]^2$ and $Y \in \{0,1\}$. Let $k^\star$ be an even integer and let $p \in (1/2, 1]$. For all $i, j \in \{1, \dots, 2^{k^\star/2}\}$, we denote C_{ij} the cell $\left[\frac{i-1}{2^{k^\star/2}}, \frac{i}{2^{k^\star/2}}\right) \times \left[\frac{j-1}{2^{k^\star/2}}, \frac{j}{2^{k^\star/2}}\right)$. We arbitrary assign a color (black or white) to each cell, which has a direct influence on the distribution of Y in the cell. More precisely, for $x \in C_{ij}$,

$$\mathbb{P}[Y = 1 | X = x] = \begin{cases} p & \text{if } C_{ij} \text{ is a black cell,} \\ 1-p & \text{if } C_{ij} \text{ is a white one.} \end{cases}$$

We define $\mathcal{B}$ (resp. $\mathcal{W}$) as the union of black (resp. white) cells and $N_{\mathcal{B}} \in \{0, \dots, 2^{k^\star}\}$ (resp. $N_{\mathcal{W}}$) as the number of black (resp. white) cells. Note that $N_{\mathcal{W}} = 2^{k^\star} - N_{\mathcal{B}}$. The location and the numbers of the black and white cells are arbitrary. This distribution corresponds to a *generalized chessboard* structure. The whole distribution is thus parameterized by $k^\star$ ($2^{k^\star}$ is the total number of cells), p and $N_{\mathcal{B}}$. Examples of this distribution are depicted in Figures 7 and 6 for different configurations.

Why such a structured setting? The data distribution introduced above is highly structured, which can be seen as a restrictive study setting. Outside this specific framework, it seems difficult for shallow tree networks to improve over a single tree. For instance, consider a more general distribution such as

$$\mathbb{P}[Y = 1 | X = x] = P_{ij} \text{ when } x \in C_{ij},$$

where P_{ij} is a random variable drawn uniformly in $[0,1]$.

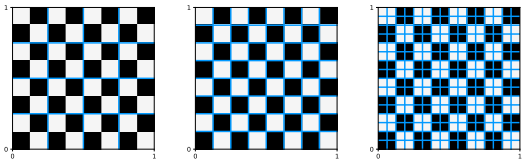
Lemma 1. *Consider the previous setting with $k \geq k^\star$. In the infinite sample setting, the risks of a single tree and a shallow tree network are given by $R(\hat{r}_{k,0,\infty}) = 0$ and*

$$R(\hat{r}_{k,1,\infty}) \geq \frac{1}{48}\left(1 - \frac{8}{2^{k^\star} - 1}\right) + \frac{1}{2^{2^{k^\star}}}\frac{9}{24}.$$

Lemma 1 highlights the fact that a tree network has a positive bias, which is not the case for a single tree. Besides, by letting $k^\star$ tend to infinity (that is the size of the cells tends to zero), the above chessboard distribution boils down to a very generic classification framework. In this latter case, the tree network performs poorly since its risk is lower bounded by $1/48$. In short, when the data distribution is disparate across the feature space, the averaging performed by the second tree leads to a biased regressor. Note that Lemma 1 involves a shallow tree network, performing only one cut on the second layer. But similar conclusions could be drawn for a deeper second-layer tree, until its depth reaches $k^\star$. Indeed, considering $\hat{r}_{k,k^\star,\infty}$ would result in an unbiased regressor, with comparable performances as of a single tree, while being much more complex.

Armed with Lemma 1, we believe that the intrinsic structure of DF and tree networks makes them useful to detect similar patterns spread across the feature space. This makes the generalized chessboard distribution particularly well suited for analyzing such behavior. The risk of a shallow tree network in the infinite sample regime for the generalized chessboard distribution is studied in Lemma 2.

Lemma 2. *Assume that the data follows the generalized chessboard distribution described above with parameter $k^\star$, $N_{\mathcal{B}}$ and p. In the infinite sample regime, the following holds for the shallow tree network $\hat{r}_{k,k',n}$ (Definition 1).*



(a) Depth 4 (b) Depth 6 (c) Depth 8

Figure 7: Chessboard data distribution for $k^\star = 6$ and $N_{\mathcal{B}} = 2^{k^\star - 1}$. Partition of the (first) encoding tree of depth 4, 6, 8 (from left to right) is displayed in blue. The optimal depth of a single centered tree for this chessboard distribution is 6.

(i) ***Shallow encoding tree.*** *Let $k < k^\star$. The risk of the shallow tree network is minimal for all configurations of the chessboard if the second-layer tree is of depth $k' \geq k^\star$ and if the $k^\star$ first cuts are performed along raw features only.*

(ii) ***Deep encoding tree.*** *Let $k \geq k^\star$. The risk of the shallow tree network is minimal for all configurations of the chessboard if the second-layer tree is of depth $k' \geq 1$ and if the first cut is performed along the new feature $\bar{Y}_{L_n(X)}$.*

In the infinite sample regime, Lemma 2 shows that the pre-processing is useless when the encoding tree is shallow ($k < k^\star$): the second tree cannot leverage on the partition of the first one and needs to build a finer partition from zero.

Lemma 2 also provides an interesting perspective on the second-layer tree which either acts as a copy of the first-layer tree or can simply be of depth one.

Remark 2. *The results established in Lemma 2 for centered-tree networks also empirically hold for CART ones (see Figures 4,S10,S13,S15,S17,S19: (i) the second-layer CART trees always make their first cut on the new feature and always near 1/2; (ii) if the first-layer CART is biased, then the second-layer tree will not improve the accuracy of the first tree (see Figure 4 (top)); (iii) if the first-layer CART is developed enough, then the second-layer CART acts as a variance reducer (see Figure 4, middle and bottom).*

4.3 Main results

Building on Lemma 1 and 2, we now focus on a shallow network whose second-layer tree is of depth one, and whose first cut is performed along the new feature $\bar{Y}_{L_n(X)}$ at 1/2. Two main regimes of training can be therefore identified when the first tree is either shallow ($k < k^\star$) or deep ($k \geq k^\star$).

In the first regime ($k < k^\star$), to establish precise non-asymptotics bounds, we study the balanced chessboard distribution (see Figure 7). Such a distribution has been studied in the unsupervised literature, in order to generate distribution for X via copula theory [10, 11] or has been mixed with other distribution in the RF framework [5]. Intuitively, this is a worst-case configuration for centered trees in terms of bias. Indeed, if $k < k^\star$, each leaf contains the same number of black and white cells. Therefore in expectation the mean value of the leaf is 1/2 which is non informative.

Proposition 3 (Risk of a single tree and a shallow tree network when $k < k^\star$)**.** *Assume that the data is drawn according to a balanced chessboard distribution with parameters $k^\star$, $N_{\mathcal{B}} = 2^{k^\star - 1}$ and $p > 1/2$ (see Figure 7).*

1. *Consider a single tree $\hat{r}_{k,0,n}$ of depth $k \in \mathbb{N}^\star$. We have,*

$$R(\hat{r}_{k,0,n}) \leq \left(p - \frac{1}{2}\right)^2 + \frac{2^k}{2(n+1)} + \frac{(1-2^{-k})^n}{4};$$

and

$$R(\hat{r}_{k,0,n}) \geq \left(p - \frac{1}{2}\right)^2 + \frac{2^k}{4(n+1)} + \frac{(1-2^{-k})^n}{4}\left(1 - \frac{2^k}{n+1}\right).$$

2. Consider the shallow tree network $\hat{r}_{k,1,n}$. We have

$$R(\hat{r}_{k,1,n}) \leq \left(p - \frac{1}{2}\right)^2 + \frac{2^{k/2+3}(p-\frac{1}{2})}{\sqrt{\pi n}} + \frac{7 \cdot 2^{2k+2}}{\pi^2(n+1)}(1+\varepsilon_{k,p}) + \frac{p^2+(1-p)^2}{2}\left(1-2^{-k}\right)^n$$

where $\varepsilon_{k,p} = o(2^{-k/2})$ uniformly in p, and

$$R(\hat{r}_{k,1,n}) \geq \left(p - \frac{1}{2}\right)^2.$$

First, note that our bounds are tight in both cases ($k < k^\star$ and $k \geq k^\star$) since the rates of the upper bounds match that of the lower ones. The first statement in Proposition 3 quantifies the bias of a single tree of depth $k < k^\star$: the term $(p-1/2)^2$ appears in both the lower and upper bounds, which means that no matter how large the training set is, the risk of the tree does not tend to zero. The shallow tree network suffers from the same bias term as soon as the first-layer tree is not deep enough. Here, the flaws of the first-layer tree transfer to the whole network. In all bounds, the term $(1-2^{-k})^n$ corresponding to the probability of X falling into an empty cell is classic and cannot be eliminated for centered trees, whose splitting strategy is independent of the dataset.

Proposition S8 in the Supplementary Materials extends the previous result to the case of a random chessboard, in which each cell has a probability of being black or white. The same phenomenon is observed: the bias of the first layer tree is not reduced, even in the infinite sample regime.

In the second regime ($k \geq k^\star$), the tree network may improve over a single tree as shown in Proposition 4.

Proposition 4 (Risk of a single tree and a shallow tree network when $k \geq k^\star$)**.** *Consider a generalized chessboard with parameters $k^\star$, $N_\mathcal{B}$ and $p > 1/2$.*

1. Consider a single tree $\hat{r}_{k,0,n}$ of depth $k \in \mathbb{N}^\star$. We have

$$R(\hat{r}_{k,0,n}) \leq \frac{2^k p(1-p)}{n+1} + \left(p^2+(1-p)^2\right)\frac{(1-2^{-k})^n}{2},$$

and

$$R(\hat{r}_{k,0,n}) \geq \frac{2^{k-1}p(1-p)}{n+1} + \left(p^2+(1-p)^2 - \frac{2^k p(1-p)}{n+1}\right)\frac{(1-2^{-k})^n}{2}.$$

2. Consider the shallow tree network $\hat{r}_{k,1,n}$. Letting

$$\bar{p}_\mathcal{B}^2 = \left(\frac{N_\mathcal{B}}{2^{k^\star}}p^2 + \frac{2^{k^\star}-N_\mathcal{B}}{2^{k^\star}}(1-p)^2\right)(1-2^{-k})^n,$$

we have

$$R(\hat{r}_{k,1,n}) \leq 2 \cdot \frac{p(1-p)}{n+1} + \frac{2^{k+1}\varepsilon_{n,k,p}}{n} + \bar{p}_\mathcal{B}^2,$$

where $\varepsilon_{n,k,p} = n(1 - \frac{1-e^{-2(p-\frac{1}{2})^2}}{2^k})^n$, and for all $n \geq 2^{k+1}(k+1)$,

$$R(\hat{r}_{k,1,n}) \geq \frac{2p(1-p)}{n} - \frac{2^{k+3}(1-\rho_{k,p})^n}{n} + \bar{p}_\mathcal{B}^2,$$

where $0 < \rho_{k,p} < 1$ depends only on p and k.

Proposition 4 shows that there exists a benefit from using this network when the first-layer tree is deep enough. In this case, the risk of the shallow tree network is $O(1/n)$ whereas that of a single tree is $O(2^k/n)$. In presence of complex and highly structured data (large $k^\star$ and similar distribution in different areas of the input space), the shallow tree network benefits from a variance reduction phenomenon by a factor 2^k. These theoretical bounds are numerically assessed in the Supplementary Materials (see Figures S33 to S38) showing their tightness for a particular choice of the chessboard configuration.

5 Conclusion

In this paper, we study both numerically and theoretically DF and its elementary components. We show that stacking layers of trees (and forests) may improve the predictive performance of the algorithm. However, most of the improvements rely on the first DF-layers. We show that the performance of a shallow tree network (composed of single CART) depends on the depth of the first-layer tree. When the first-layer tree is deep enough, the second-layer tree may build upon the new features created by the first tree by acting as a variance reducer.

To quantify this phenomenon, we propose a first theoretical analysis of a shallow tree network (composed of centered trees). Our study exhibits the crucial role of the first (encoding) layer: if the first-layer tree is biased, then the entire shallow network inherits this bias, otherwise the second-layer tree acts as a good variance reducer. One should note that this variance reduction cannot be obtained by averaging many trees, as in RF structure: the variance of an averaging of centered trees with depth k is of the same order as one of these individual trees [4, 15], whereas two trees in cascade (the first one of depth k and the second of depth 1) may lead to a variance reduction by a 2^k factor. This highlights the benefit of tree-layer architectures over standard ensemble methods. We thus believe that this first theoretical study of this shallow tree network paves the way of the mathematical understanding of DF.

First-layer trees, and more generally the first layers in DF architecture, can be seen as data-driven encoders. More precisely, the first layers in DF create an automatic embedding of the data, building on the specific conditional relation between the output and the inputs, therefore potentially improving the performance of the overall structure. Since preprocessing is nowadays an important part of all machine learning pipelines, we believe that our analysis is interesting beyond the framework of DF.

References

[1] J. S. Bergstra, R. Bardenet, Y. Bengio, and K. Balázs. Algorithms for hyper-parameter optimization. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 2546–2554. Curran Associates, Inc., 2011.

[2] A. Berrouachedi, R. Jaziri, and G. Bernard. Deep cascade of extra trees. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 117–129. Springer, 2019.

[3] A. Berrouachedi, R. Jaziri, and G. Bernard. Deep extremely randomized trees. In *International Conference on Neural Information Processing*, pages 717–729. Springer, 2019.

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

[5] G. Biau, L. Devroye, and G. Lugosi. Consistency of random forests and other averaging classifiers. *Journal of Machine Learning Research*, 9(Sep):2015–2033, 2008.

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

[7] F. Cribari-Neto, N. L. Garcia, and K. LP Vasconcellos. A note on inverse moments of binomial variates. *Brazilian Review of Econometrics*, 20(2):269–277, 2000.

[8] Wei Fan, Haixun Wang, Philip S Yu, and Sheng Ma. Is random model better? on its accuracy and efficiency. In *Third IEEE International Conference on Data Mining*, pages 51–58. IEEE, 2003.

[9] J. Feng, Y. Yu, and Z-H Zhou. Multi-layered gradient boosting decision trees. In *Advances in neural information processing systems*, pages 3551–3561, 2018.

[10] Soumyadip Ghosh and Shane G Henderson. Chessboard distributions and random vectors with specified marginals and covariance matrix. *Operations Research*, 50(5):820–834, 2002.

[11] Soumyadip Ghosh and Shane G Henderson. Patchwork distributions. In *Advancing the Frontiers of Simulation*, pages 65–86. Springer, 2009.

[12] Y. Guo, S. Liu, Z. Li, and X. Shang. Bcdforest: a boosting cascade deep forest model towards the classification of cancer subtypes based on gene expression data. *BMC bioinformatics*, 19(5):118, 2018.

[13] M. Jeong, J. Nam, and B. C. Ko. Lightweight multilayer random forests for monitoring driver emotional status. *IEEE Access*, 8:60344–60354, 2020.

[14] S. Kim, M. Jeong, and B. C. Ko. Interpretation and simplification of deep forest. *arXiv preprint arXiv:2001.04721*, 2020.

[15] J. M. Klusowski. Sharp analysis of a simple model for random forests. *arXiv preprint arXiv:1805.02587*, 2018.

[16] B. Liu, W. Guo, X. Chen, K. Gao, X. Zuo, R. Wang, and A. Yu. Morphological attribute profile cube and deep random forest for small sample classification of hyperspectral image. *IEEE Access*, 8:117096–117108, 2020.

[17] D. A. Melis and T. Jaakkola. Towards robust interpretability with self-explaining neural networks. In *Advances in Neural Information Processing Systems*, pages 7775–7784, 2018.

[18] K. Miller, C. Hettinger, J. Humpherys, T. Jarvis, and D. Kartchner. Forward thinking: Building deep random forests. *arXiv*, 2017.

[19] M. Pang, K. Ting, P. Zhao, and Z. Zhou. Improving deep forest by confidence screening. In *2018 IEEE International Conference on Data Mining (ICDM)*, pages 1194–1199, 2018.

[20] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[21] R. Su, X. Liu, L. Wei, and Q. Zou. Deep-resp-forest: A deep forest model to predict anti-cancer drug response. *Methods*, 166:91–102, 2019.

[22] L. Sun, Z. Mo, F. Yan, L. Xia, F. Shan, Z. Ding, B. Song, W. Gao, W. Shao, F. Shi, H. Yuan, H. Jiang, D. Wu, Y. Wei, Y. Gao, H. Sui, D. Zhang, and D. Shen. Adaptive feature selection guided deep forest for covid-19 classification with chest ct. *IEEE Journal of Biomedical and Health Informatics*, 24(10):2798–2805, 2020.

[23] L. V Utkin and M. A Ryabinin. Discriminative metric learning with deep forest. *arXiv preprint arXiv:1705.09620*, 2017.

[24] L. V Utkin and K. D Zhuk. Improvement of the deep forest classifier by a set of neural networks. *Informatica*, 44(1), 2020.

[25] X. Zeng, S. Zhu, Y. Hou, P. Zhang, L. Li, J. Li, L F. Huang, S. J Lewis, R. Nussinov, and F. Cheng. Network-based prediction of drug–target interactions using an arbitrary-order proximity embedded deep forest. *Bioinformatics*, 36(9):2805–2812, 2020.

[26] Y. Zhang, J. Zhou, W. Zheng, J. Feng, L. Li, Z. Liu, M. Li, Z. Zhang, C. Chen, X. Li, et al. Distributed deep forest and its application to automatic detection of cash-out fraud. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(5):1–19, 2019.

[27] S. Zheng, Y. Song, T. Leung, and I. Goodfellow. Improving the robustness of deep neural networks via stability training. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 4480–4488, 2016.

[28] Z Zhou and J. Feng. Deep forest: Towards an alternative to deep neural networks. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 3553–3559, 2017.

[29] Z. Zhou and J. Feng. Deep forest. *National Science Review*, 6(1):74–86, 2019.

S1 Additional figures

S1.1 Computation times for Section 3

	Yeast	Housing	Letter	Adult	Airbnb	Higgs
Default DF time	13m19s	9m38s	20m31	13m57s	23m23s	43m53s
Light DF time	7s	6s	8s	8s	10s	13s
Default DF MC (MB)	11	6	174	139	166	531
Light DF MC (MB)	5	4	109	72	100	318

Table S2: Comparing the time and memory consumption of DF and Light DF.

S1.2 Table of best configurations, supplementary to Section ??

Dataset	Best configuration hyperparam.	Mean optimal sub-model (10 tries)
Adult	6 forests, 20 trees, max depth 30	1.0
Higgs	10 forests, 800 trees, max depth 30	5.1
Letter	8 forests, 500 trees, max depth None (default)	1.0
Yeast	6 forests, 280 trees, max depth 30	2.1
Airbnb	4 forests, 150 trees, max depth 30	2.0
Housing	10 forests, 280 trees, max depth 10	11.2

Table S3: Details of the best configurations obtained in Figures 2 and 3.

S1.3 Additional figures to Section 3.2

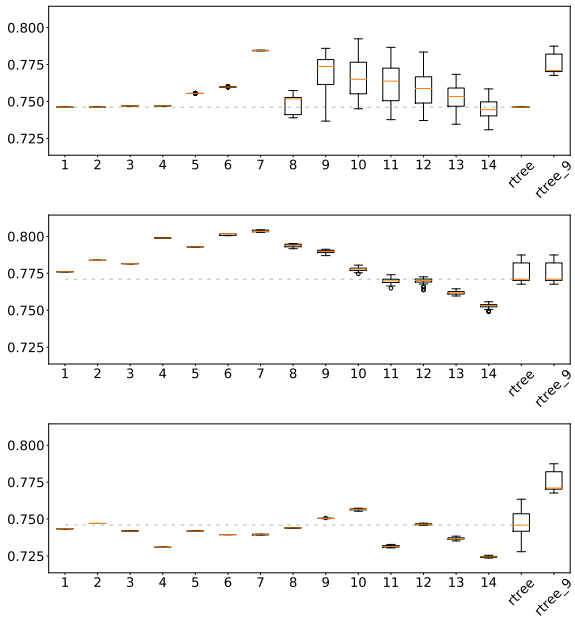


Figure S8: Adult dataset. Accuracy of a two-layer tree architecture w.r.t. the second-layer tree depth, when the first-layer (encoding) tree is of depth 2 (top), 9 (middle), and 15 (bottom). `rtree` is a single tree of respective depth 2 (top), 9 (middle), and 15 (bottom), applied on raw data. For this dataset, the optimal depth of a single tree is 9 and the tree with the optimal depth is depicted as `rtree_9` in each plot. The green dashed line indicates the median score of the `rtree`. All boxplots are obtained by 10 different runs.

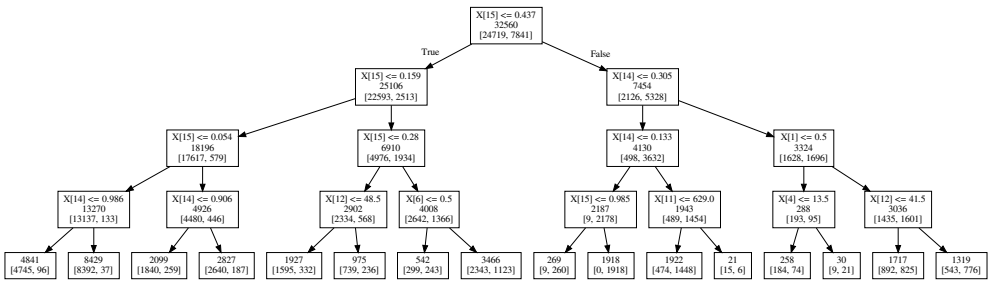


Figure S9: Adult dataset. Second-layer tree structure of depth 4 when the first-layer tree is of depth 9 (optimal depth). Raw features range from X[0] to X[13], X[14] and X[15] are the features built by the first-layer tree.

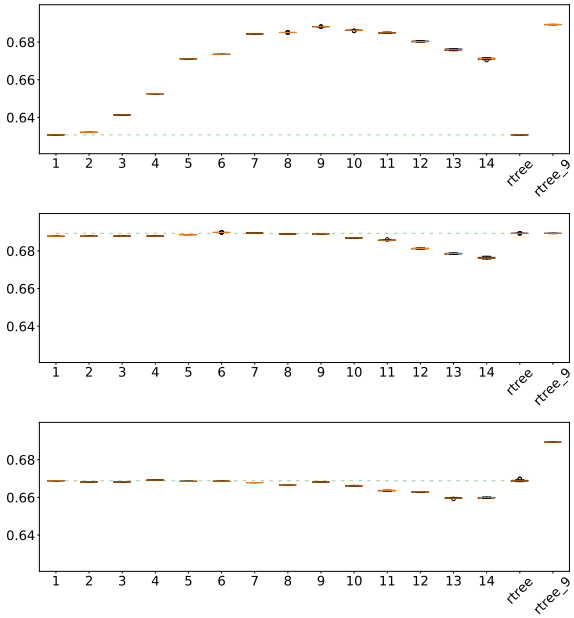


Figure S10: Higgs dataset. Accuracy of a two-layer tree architecture w.r.t. the second-layer tree depth, when the first-layer (encoding) tree is of depth 2 (top), 9 (middle), and 15 (bottom). `rtree` is a single tree of respective depth 2 (top), 9 (middle), and 15 (bottom), applied on raw data. For this dataset, the optimal depth of a single tree is 9 and the tree with the optimal depth is depicted as `rtree_9` in each plot. The green dashed line indicates the median score of the `rtree`. All boxplots are obtained by 10 different runs.

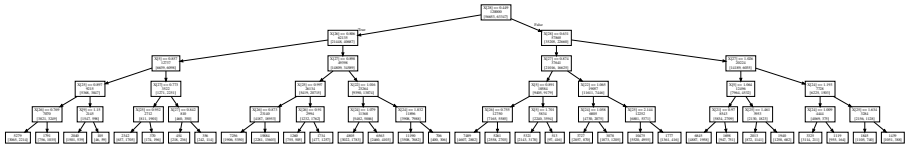


Figure S11: Higgs dataset. Second-layer tree structure of depth 5 when the first-layer tree is of depth 2 (low depth). Raw features range from X[0] to X[13], X[14] and X[15] are the features built by the first-layer tree.

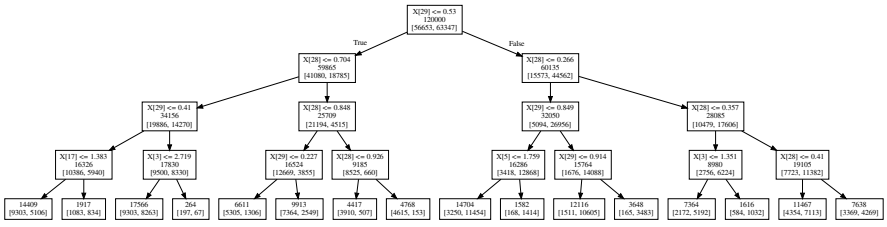


Figure S12: Higgs dataset. Second-layer tree structure of depth 4 when the first-layer tree is of depth 9 (optimal depth). Raw features range from X[0] to X[27], X[28] and X[29] are the features built by the first-layer tree.

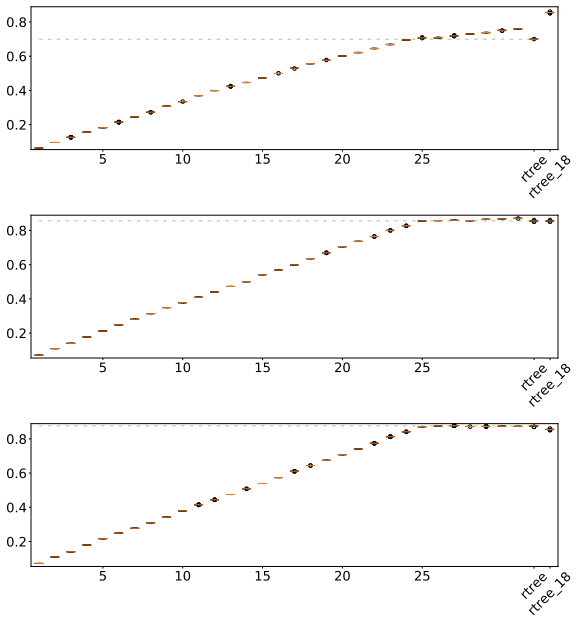


Figure S13: Letter dataset. Accuracy of a two-layer tree architecture w.r.t. the second-layer tree depth, when the first-layer (encoding) tree is of depth 10 (top), 18 (middle), and 26 (bottom). `rtree` is a single tree of respective depth 10 (top), 18 (middle), and 26 (bottom), applied on raw data. For this dataset, the optimal depth of a single tree is 18 and the tree with the optimal depth is depicted as `rtree_18` in each plot. The green dashed line indicates the median score of the `rtree`. All boxplots are obtained by 10 different runs.

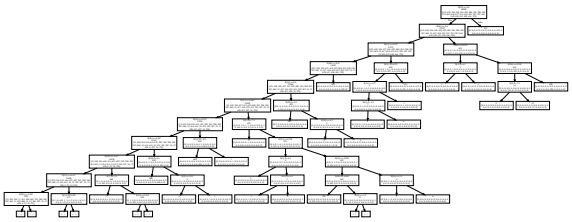


Figure S14: Letter dataset. Second-layer tree structure of depth 30 when the first-layer tree is of depth 18 (optimal depth). We only show the first part of the tree up to depth 10. Raw features range from X[0] to X[15]. The features built by the first-layer tree range from X[16] to X[41].

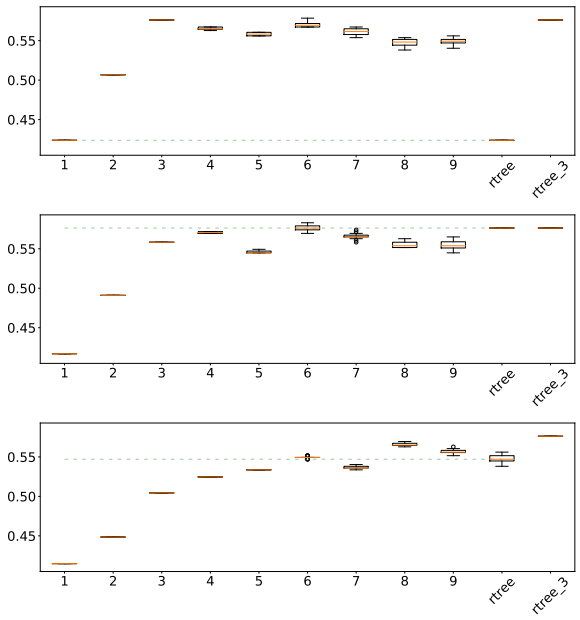


Figure S15: Yeast dataset. Accuracy of a two-layer tree architecture w.r.t. the second-layer tree depth, when the first-layer (encoding) tree is of depth 1 (top), 3 (middle), and 8 (bottom). `rtree` is a single tree of respective depth 1 (top), 3 (middle), and 8 (bottom), applied on raw data. For this dataset, the optimal depth of a single tree is 3 and the tree with the optimal depth is depicted as `rtree_3` in each plot. The green dashed line indicates the median score of the `rtree`. All boxplots are obtained by 10 different runs.

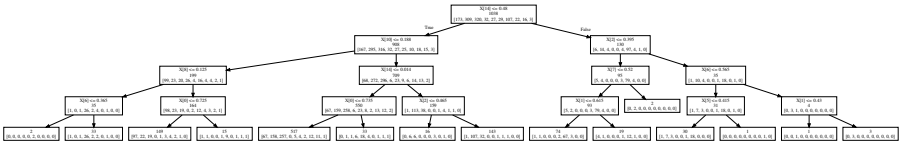


Figure S16: Yeast dataset. Second-layer tree structure of depth 4 when the first-layer tree is of depth 3 (optimal depth). Raw features range from X[0] to X[7]. The features built by the first-layer tree range from X[8] to X[17].

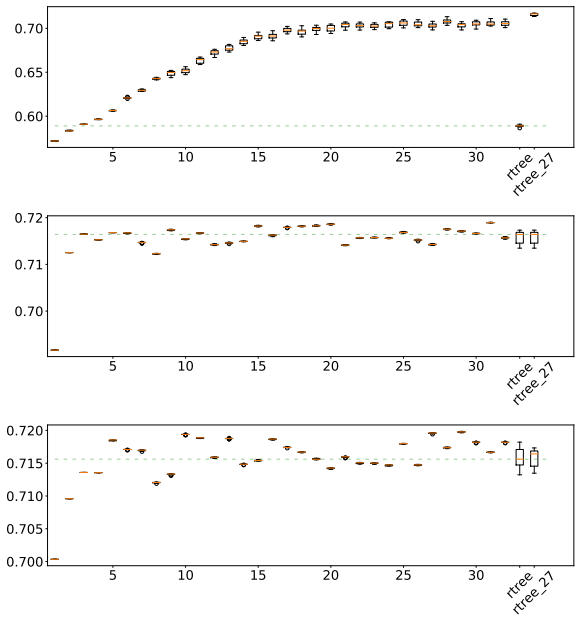


Figure S17: Airbnb dataset. R^2 score of a two-layer tree architecture w.r.t. the second-layer tree depth, when the first-layer (encoding) tree is of depth 10 (top), 27 (middle), and 32 (bottom). `rtree` is a single tree of respective depth 10 (top), 27 (middle), and 32 (bottom), applied on raw data. For this dataset, the optimal depth of a single tree is 27 and the tree with the optimal depth is depicted as `rtree_27` in each plot. The green dashed line indicates the median score of the `rtree`. All boxplots are obtained by 10 different runs.

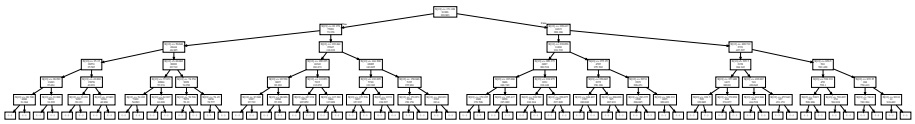


Figure S18: Airbnb dataset. Second-layer tree structure of depth 28 when the first-layer tree is of depth 26 (optimal depth). We only show the first part of the tree up to depth 5. Raw features range from X[0] to X[12], X[13] is the feature built by the first-layer tree.

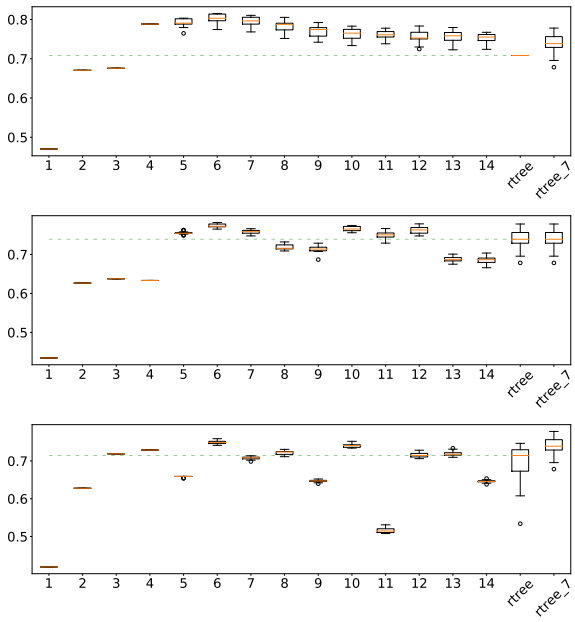


Figure S19: Housing dataset. R^2 score of a two-layer tree architecture w.r.t. the second-layer tree depth, when the first-layer (encoding) tree is of depth 3 (top), 7 (middle), and 12 (bottom). `rtree` is a single tree of respective depth 3 (top), 7 (middle), and 12 (bottom), applied on raw data. For this dataset, the optimal depth of a single tree is 9 and the tree with the optimal depth is depicted as `rtree_7` in each plot. The green dashed line indicates the median score of the `rtree`. All boxplots are obtained by 10 different runs.

Figure S20: Housing dataset. Second-layer tree structure of depth 10 when the first-layer tree is of depth 7 (optimal depth). We only show the first part of the tree up to depth 5. Raw features range from X[0] to X[60], X[61] is the feature built by the first-layer tree.

S1.4 Additional figures to Section ??

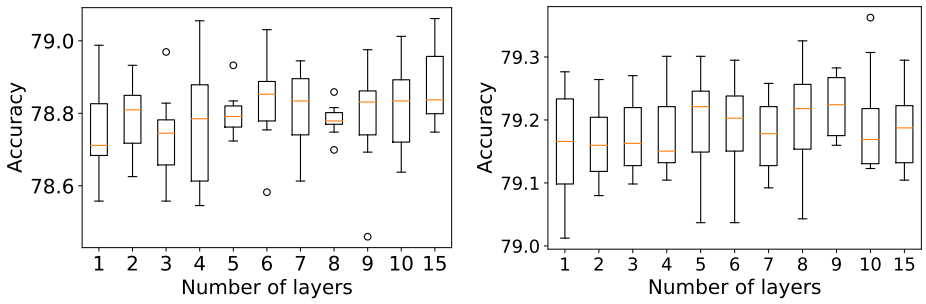


Figure S21: Adult dataset. Boxplots over 10 tries of the accuracy of a DF with 1 forest by layer (left) or 4 forests by layer (right), with respect to the DF maximal number of layers.

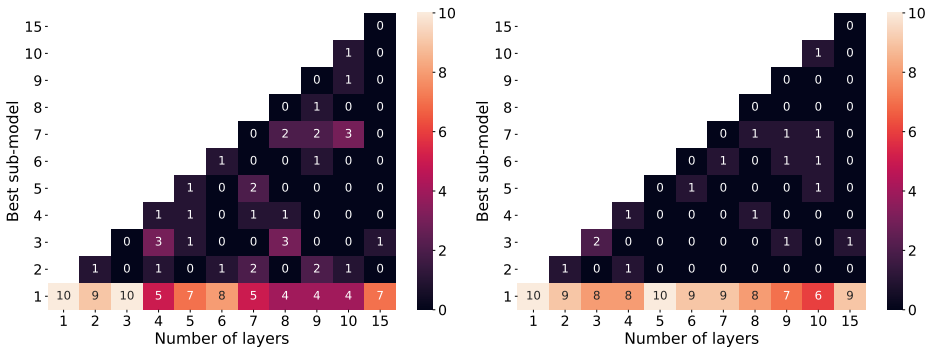


Figure S22: Adult dataset. Heatmap counting the index of the sub-optimal model over 10 tries of a default DF with 1 (Breiman) forest per layer (left) or 4 forests (2 Breiman, 2 CRF) per layer (right), with respect to the maximal number of layers.

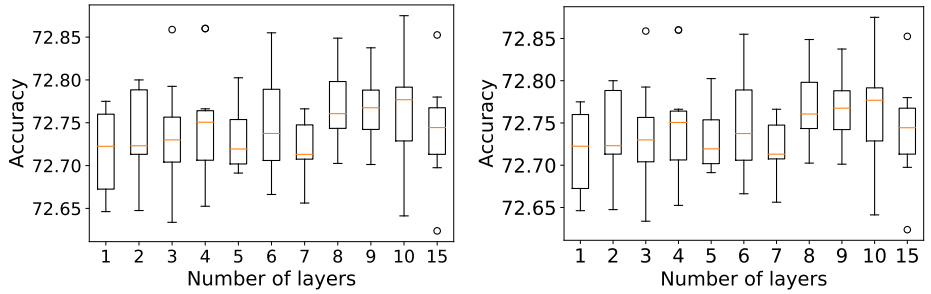


Figure S23: Higgs dataset. Boxplots over 10 tries of the accuracy of a DF with 1 forest by layer (left) or 4 forests by layer (right), with respect to the DF maximal number of layers.

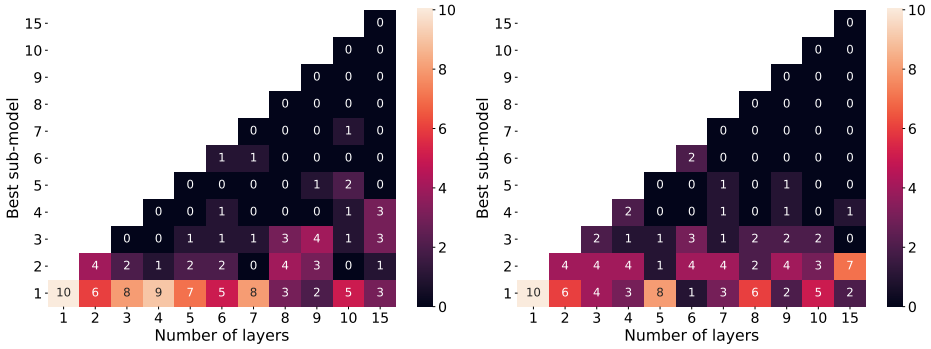


Figure S24: Higgs dataset. Heatmap counting the index of the sub-optimal model over 10 tries of a default DF with 1 (Breiman) forest per layer (left) or 4 forests (2 Breiman, 2 CRF) per layer (right), with respect to the maximal number of layers.

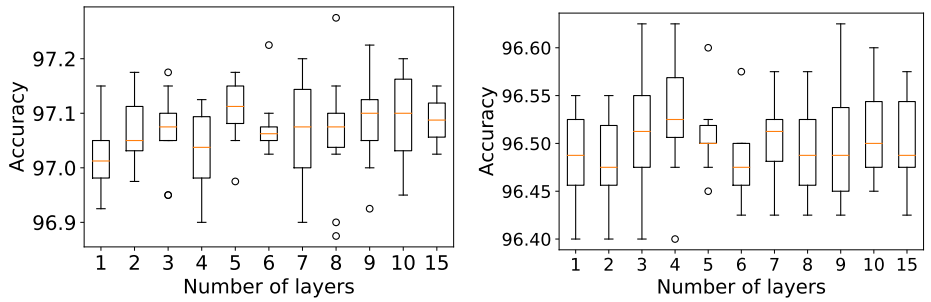


Figure S25: Letter dataset. Boxplots over 10 tries of the accuracy of a DF with 1 forest by layer (left) or 4 forests by layer (right), with respect to the DF maximal number of layers.

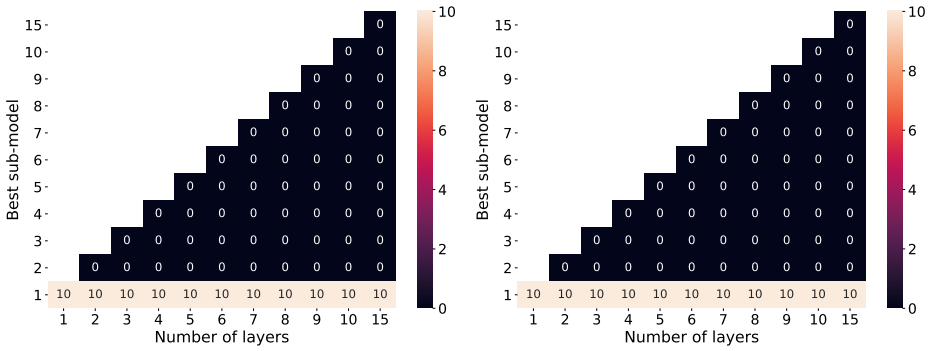


Figure S26: Letter dataset. Heatmap counting the index of the sub-optimal model over 10 tries of a default DF with 1 (Breiman) forest per layer (left) or 4 forests (2 Breiman, 2 CRF) per layer (right), with respect to the maximal number of layers.

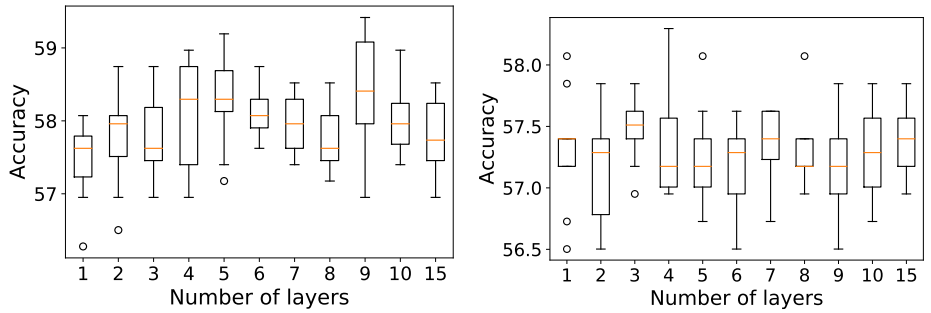


Figure S27: Yeast dataset. Boxplots over 10 tries of the accuracy of a DF with 1 forest by layer (left) or 4 forests by layer (right), with respect to the DF maximal number of layers.

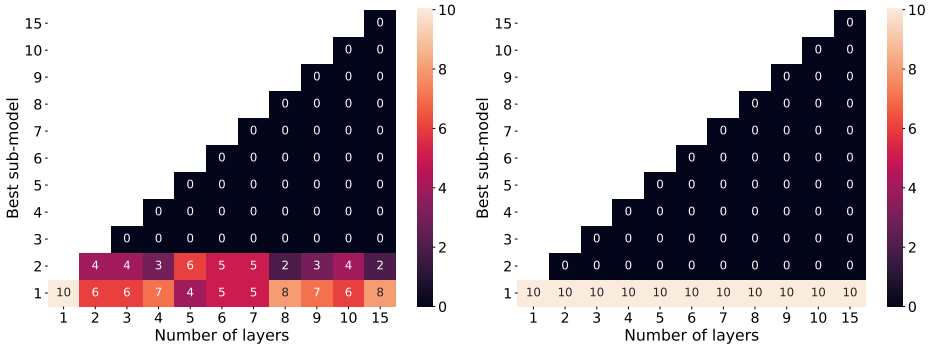


Figure S28: Yeast dataset. Heatmap counting the index of the sub-optimal model over 10 tries of a default DF with 1 (Breiman) forest per layer (left) or 4 forests (2 Breiman, 2 CRF) per layer (right), with respect to the maximal number of layers.

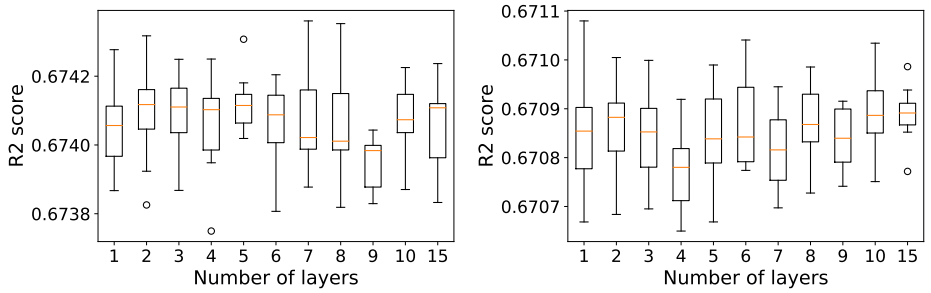


Figure S29: Airbnb dataset. Boxplots over 10 tries of the accuracy of a DF with 1 forest by layer (left) or 4 forests by layer (right), with respect to the DF maximal number of layers.

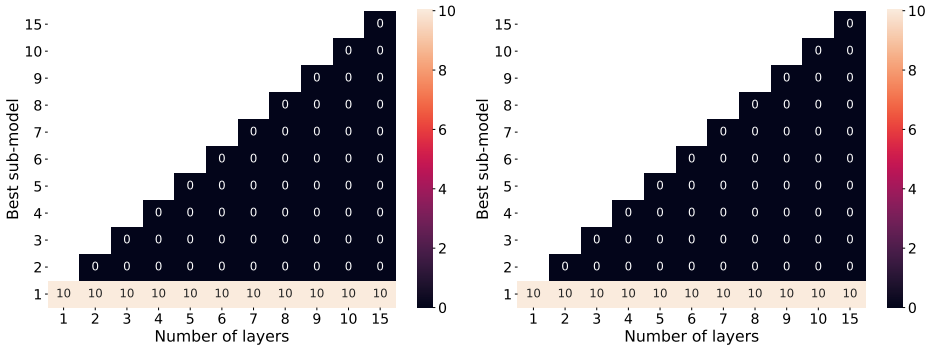


Figure S30: Airbnb datase. Heatmap counting the index of the sub-optimal model over 10 tries of a default DF with 1 (Breiman) forest per layer (left) or 4 forests (2 Breiman, 2 CRF) per layer (right), with respect to the maximal number of layers.

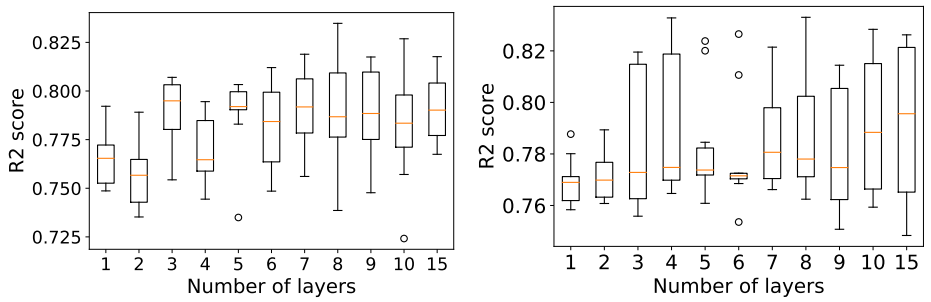


Figure S31: Housing dataset. Boxplots over 10 tries of the accuracy of a DF with 1 forest by layer (left) or 4 forests by layer (right), with respect to the DF maximal number of layers.

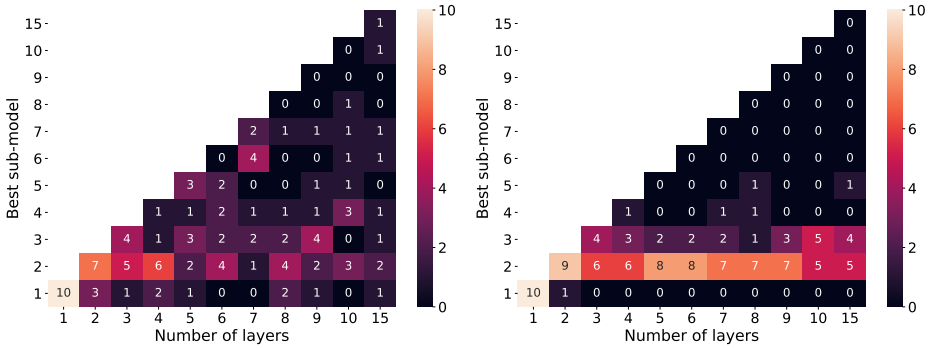


Figure S32: Housing dataset. Heatmap counting the index of the sub-optimal model over 10 tries of a default DF with 1 (Breiman) forest per layer (left) or 4 forests (2 Breiman, 2 CRF) per layer (right), with respect to the maximal number of layers.

S1.5 Additional figures to Section 4

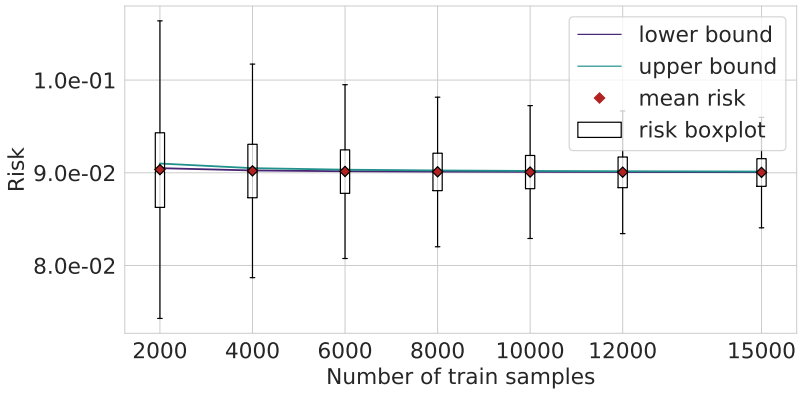


Figure S33: Illustration of the theoretical bounds for a single tree of Proposition 3 1. for a chessboard with parameters $k^\star = 4$, $N_{\mathcal{B}} = 2^{k^\star - 1}$, and $p = 0.8$. The single tree is of depth $k = 2$. We draw a sample of size n (x-axis), and a single tree $r_{k,0,n}$ is fitted for which the theoretical risk is evaluated. Each boxplot is built out of 20 000 repetitions. The outliers are not shown for the sake of presentation.

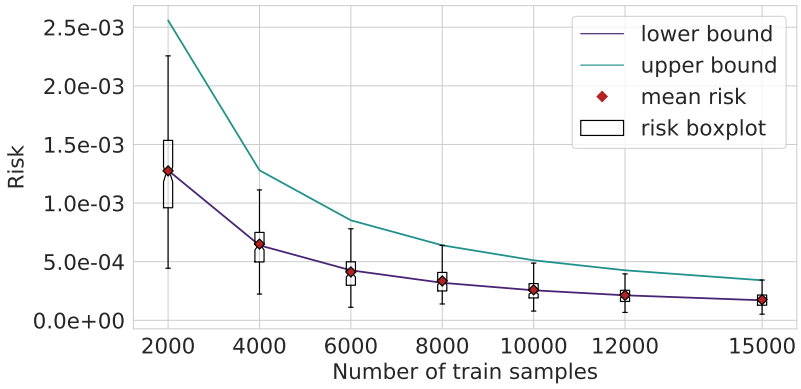


Figure S34: Illustration of the theoretical bounds for a single tree of Proposition 4 1. for a chessboard with parameters $k^\star = 4$, $N_{\mathcal{B}} = 2^{k^\star - 1}$ and $p = 0.8$. The single tree is of depth $k = 4$. We draw a sample of size n (x-axis), and a single tree $r_{k,0,n}$ is fitted for which the theoretical risk is evaluated. Each boxplot is built out of 20 000 repetitions. The outliers are not shown for the sake of presentation.

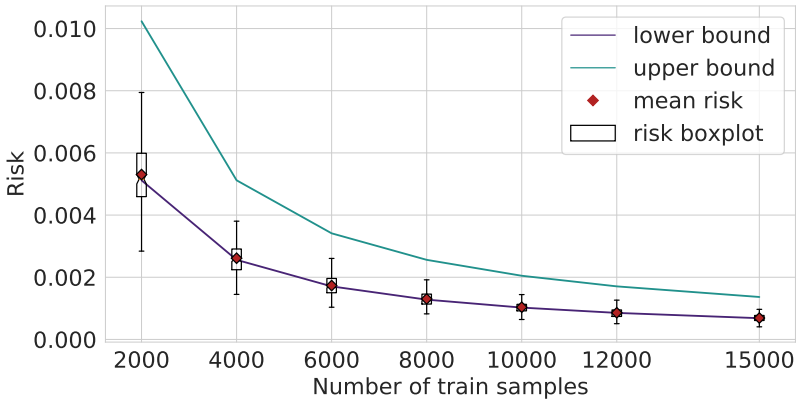


Figure S35: Illustration of the theoretical bounds for a single tree of Proposition 4 1. for a chessboard with parameters $k^\star = 4$, $N_{\mathcal{B}} = 2^{k^\star - 1}$ and $p = 0.8$. The single tree is of depth $k = 6$. We draw a sample of size n (x-axis), and a single tree $r_{k,0,n}$ is fitted for which the theoretical risk is evaluated. Each boxplot is built out of 20 000 repetitions. The outliers are not shown for the sake of presentation.

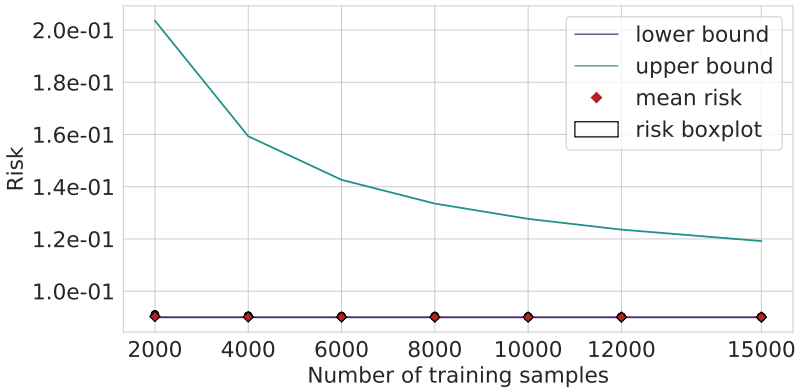


Figure S36: Illustration of the theoretical bounds for a shallow tree network of Proposition 3 2. for a chessboard with parameters $k^\star = 4$, $N_{\mathcal{B}} = 2^{k^\star - 1}$ and $p = 0.8$. The first-layer tree is of depth $k = 2$. We draw a sample of size n (x-axis), and a single tree $r_{k,0,n}$ is fitted for which the theoretical risk is evaluated. Each boxplot is built out of 20 000 repetitions. The outliers are not shown for the sake of presentation. Note that in such a case, the theoretical lower bound is constant and equal to the bias term.

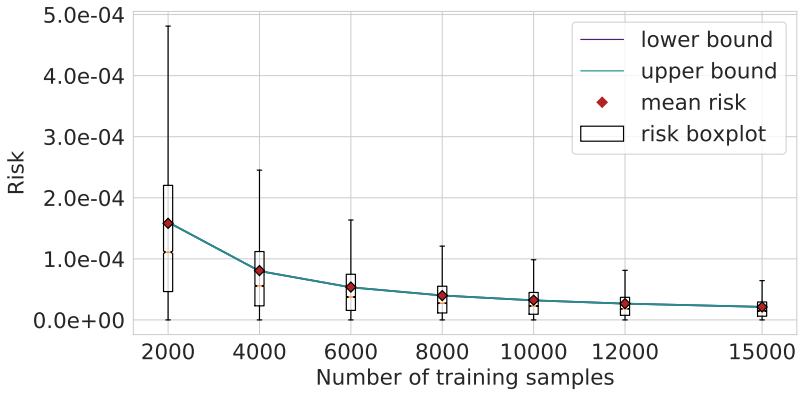


Figure S37: Illustration of the theoretical bounds for a shallow tree network of Proposition 4 2. for a chessboard with parameters $k^\star = 4$, $N_{\mathcal{B}} = 2^{k^\star - 1}$ and $p = 0.8$. The first-layer tree is of depth $k = 4$. We draw a sample of size n (x-axis), and a single tree $r_{k,0,n}$ is fitted for which the theoretical risk is evaluated. Each boxplot is built out of 20 000 repetitions. The outliers are not shown for the sake of presentation. Note that in such a case, the theoretical lower bound is constant and equal to the bias term. Note that the lower bound and the upper bound are merged.

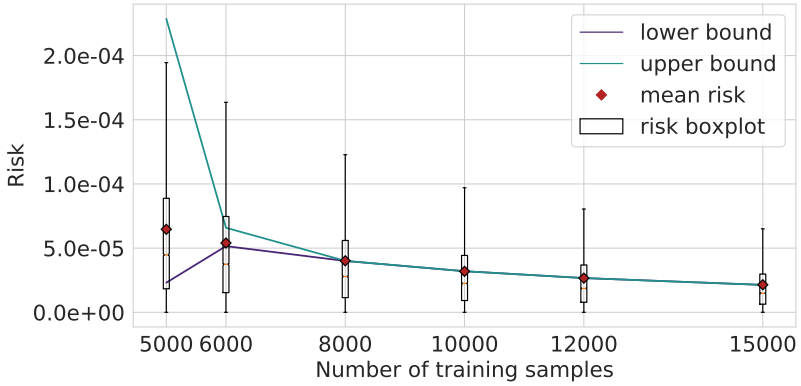


Figure S38: Illustration of the theoretical bounds for a shallow tree network of Proposition 4 2. for a chessboard with parameters $k^\star = 4$, $N_{\mathcal{B}} = 2^{k^\star - 1}$ and $p = 0.8$. The first-layer tree is of depth $k = 6$. We draw a sample of size n (x-axis), and a single tree $r_{k,0,n}$ is fitted for which the theoretical risk is evaluated. Each boxplot is built out of 20 000 repetitions. The outliers are not shown for the sake of presentation. Note that in such a case, the theoretical lower bound is constant and equal to the bias term.

S2 Technical results on binomial random variables

Lemma S5. *Let Z be a binomial $\mathfrak{B}(n,p)$, $p \in (0,1], n > 0$. Then,*

(i)
$$\frac{1-(1-p)^n}{(n+1)p} \leq \mathbb{E}\left[\frac{\mathbb{1}_{Z>0}}{Z}\right] \leq \frac{2}{(n+1)p}$$

(ii)
$$\mathbb{E}\left[\frac{1}{1+Z}\right] \leq \frac{1}{(n+1)p}$$

(iii)
$$\mathbb{E}\left[\frac{1}{1+Z^2}\right] \leq \frac{3}{(n+1)(n+2)p^2}$$

(iv)
$$\mathbb{E}\left[\frac{\mathbb{1}_{Z>0}}{\sqrt{Z}}\right] \leq \frac{2}{\sqrt{np}}$$

(v) Let k be an integer $\leq n$. Then,
$$\mathbb{E}\left[Z \mid Z \geq k\right] = np + (1-p)k\frac{\mathbb{P}\left(Z=k\right)}{\sum_{i=k}^{n}\mathbb{P}\left(Z=i\right)}.$$

(vi) Let Z be a binomial $\mathfrak{B}(n,\frac{1}{2})$, $n > 0$. Then,
$$\mathbb{E}\left[Z \mid Z \leq \lfloor\frac{n+1}{2}\rfloor - 1\right] \geq \frac{n}{2} - \left(\frac{\sqrt{n}}{\sqrt{\pi}} + \frac{2\sqrt{2n}}{\pi\sqrt{2n+1}}\right).$$

(vii) Let Z be a binomial $\mathfrak{B}(n,\frac{1}{2})$, $n > 0$. Then,
$$\mathbb{E}\left[Z \mid Z \geq \lfloor\frac{n+1}{2}\rfloor\right] \leq \frac{n}{2} + 1 + \frac{1}{\sqrt{\pi(n+1)}}.$$

Proof. The reader may refer to the Lemma 11 of [4] to see the proof of (ii), (iii) and the right-hand side of (i). The left-hand side inequality of (i) can be found in the Section 1 of [7].

(iv) The first two inequalities rely on simple analysis :

$$\begin{aligned}\mathbb{E}\left[\frac{\mathbb{1}_{Z>0}}{\sqrt{Z}}\right] &\leq \mathbb{E}\left[\frac{2}{1+\sqrt{Z}}\right] \\ &\leq \mathbb{E}\left[\frac{2}{\sqrt{1+Z}}\right].\end{aligned}$$

To go on, we adapt a transformation from Section 2 of [7] to our setting:

$$\begin{aligned}\mathbb{E}\left[\frac{2}{\sqrt{1+Z}}\right] &= \frac{2}{\Gamma(1/2)}\int_0^\infty \frac{e^{-t}}{\sqrt{t}}\mathbb{E}\left[e^{-tZ}\right]dt \\ &= \frac{2}{\Gamma(1/2)}\int_0^\infty \frac{e^{-t}}{\sqrt{t}}(1-p+pe^{-t})^n dt \\ &= \frac{2}{\Gamma(1/2)}\int_0^{-\log(1-p)} g(r)e^{-rn}dr,\end{aligned}$$

with $g(r) := p^{-1}e^{-r}\left(-\log(1+\frac{1-e^{-r}}{p})\right)^{-1/2}$ after the change of variable $(1-p+pe^{-t}) = e^{-r}$.

Let's prove that

$$g(r) \leq \frac{1}{\sqrt{rp}}. \tag{1}$$

It holds that $\log(1+x) \leq \frac{2x}{2+x}$ when $-1 < x \leq 0$, therefore

$$g(r)^2 = p^{-2}e^{-2r}\left(-\log(1+\frac{1-e^{-r}}{p})\right)^{-1} \leq p^{-2}e^{-2r}\frac{2p+e^{-r}-1}{2(1-e^{-r})}.$$

Furthermore,

$$\begin{aligned} 2p &\geq 2p\left(e^{-r}+re^{-2r}\right) \\ &\geq 2p\left(e^{-r}+re^{-2r}\right) + r\left(e^{-3r}-e^{-2r}\right) \\ &= re^{-2r}(2p-1+e^{-r}) + 2pe^{-r}, \end{aligned}$$

and then dividing by rp^2,

$$\frac{2}{rp}(1-e^{-r}) \geq \frac{1}{p^2}e^{-2r}(2p-1+e^{-r}) \qquad \iff \qquad \frac{1}{rp} \geq p^{-2}e^{-2r}\frac{2p+e^{-r}-1}{2(1-e^{-r})},$$

which proves (1).

Equation (1) leads to

$$\mathbb{E}\left[\frac{2}{\sqrt{1+Z}}\right] \leq \frac{2}{\Gamma(1/2)}\int_0^{-\log(1-p)} \frac{1}{\sqrt{pr}}e^{-rn}dr. \tag{2}$$

Note that $\Gamma(1/2) = \sqrt{\pi}$. After the change of variable $u = \sqrt{rn}$, we obtain :

$$\mathbb{E}\left[\frac{2}{\sqrt{1+Z}}\right] \leq \frac{4}{\sqrt{np\pi}}\int_0^{\sqrt{-n\log(1-p)}} e^{-u^2}du \leq \frac{4}{\sqrt{np\pi}}\int_0^{\infty} e^{-u^2}du \leq \frac{2}{\sqrt{np}}$$

which ends the proof of (iv).

(v).(a) We recall that $p = 1/2$. An explicit computation of the expectation yields :

$$\begin{aligned} \mathbb{E}\left[Z \mid Z < \lfloor\frac{n+1}{2}\rfloor\right] &= \frac{1}{\mathbb{P}\left(Z \leq \lfloor\frac{n+1}{2}\rfloor - 1\right)} \sum_{i=1}^{\lfloor\frac{n+1}{2}\rfloor-1} \frac{i}{2^n}\binom{n}{i} \\ &= \frac{2}{1}\frac{n}{2^n}\left(\frac{2^n}{2} - \frac{1}{2}\binom{n-1}{\frac{n-1}{2}}\right)\mathbb{1}_{n\%2=1} \\ &+ \frac{n}{\frac{1}{2}-\frac{1}{2}\mathbb{P}\left(Z=n/2\right)}\left(\sum_{i=1}^{n/2} i\binom{n}{i} - \frac{n}{2}\binom{n}{n/2}\right)\frac{\mathbb{1}_{n\%2=0}}{2^n} \\ &= n\left(\frac{1}{2} - \frac{1}{2^n}\binom{n-1}{\frac{n-1}{2}}\right)\mathbb{1}_{n\%2=1} + \frac{n\cdot\mathbb{1}_{n\%2=0}}{1-\mathbb{P}\left(Z=n/2\right)}\left(\frac{1}{2} - \frac{1}{2^n}\binom{n}{n/2}\right). \end{aligned}$$

We use that for all $m \in 2\mathbb{N}^*$,

$$\binom{m}{m/2} \leq \frac{2^m}{\sqrt{\pi(m/2+1/4)}} \tag{3}$$

and

$$\frac{1}{1-\mathbb{P}\left(Z=m/2\right)} \geq 1 + \frac{\sqrt{2}}{\sqrt{\pi n}}$$

where the last inequality can be obtained via a series expansion at $n = \infty$. Replacing the terms by their bounds, we have :

$$\mathbb{E}\left[Z \mid Z < \lfloor\frac{n+1}{2}\rfloor\right] \geq n\left(\left(\frac{1}{2} - \frac{1}{\sqrt{\pi(2m-1)}}\right)\mathbb{1}_{n\%2=1} + \left(1+\frac{\sqrt{2}}{\sqrt{\pi n}}\right)\left(\frac{1}{2} - \frac{2}{\sqrt{\pi(2n+1)}}\right)\mathbb{1}_{n\%2=0}\right)$$
$$\geq n\left(\frac{1}{2} - \frac{1}{\sqrt{n\pi}} - \frac{2\sqrt{2}}{\pi\sqrt{n(2n+1)}}\right)$$
$$\geq \frac{n}{2} + \sqrt{n}\left(\frac{1}{\sqrt{\pi}} - \frac{2\sqrt{2}}{\pi}\sqrt{(2n+1)}\right)$$

which ends the proof of this item (v)(a).

(v).(b) We also begin with an explicit computation of the expectation :

$$\mathbb{E}\left[Z \mid Z \geq \lfloor\frac{n+1}{2}\rfloor\right] = \frac{1}{\mathbb{P}\left(Z \geq \lfloor\frac{n+1}{2}\rfloor\right)} \sum_{i=\lfloor\frac{n+1}{2}\rfloor}^{n} \frac{i}{2^n}\binom{n}{i}$$
$$= \frac{2}{1}\frac{1}{2^n}\left(2^{n-2} + 2^{n-1} + \frac{1}{2}\binom{n-1}{\frac{n-1}{2}}\right)\mathbb{1}_{n\%2=1}$$
$$+ \frac{n}{\frac{1}{2}+\frac{1}{2}\mathbb{P}\left(Z=n/2\right)}\left(\sum_{i=\lfloor\frac{n+1}{2}\rfloor}^{n} i\binom{n}{i}\right)\frac{\mathbb{1}_{n\%2=0}}{2^n}$$
$$= \left(\frac{n}{2} + 1 + \frac{1}{2^n}\binom{n-1}{\frac{n-1}{2}}\right)\mathbb{1}_{n\%2=1} + \frac{n\cdot\mathbb{1}_{n\%2=0}}{1+\mathbb{P}\left(Z=n/2\right)}\left(\frac{1}{2} + \frac{1}{2^n}\binom{n}{n/2}\right).$$

The computation of the upper bound relies on the following inequalities : $\forall m \in 2\mathbb{N}^*$,

$$\binom{2m}{m} \leq \frac{2^{2m}}{\sqrt{\pi(m+1/4)}} \tag{4}$$

as well as

$$\frac{1}{1+\mathbb{P}\left(Z=n/2\right)} \leq 1 - \frac{\sqrt{2}}{\sqrt{\pi n}} + \frac{2}{\pi n}$$

where the last bound can be found via a series expansion at $n=\infty$. Replacing all terms by their bound and simplifying roughly gives the result. □

Lemma S6 (Uniform Bernoulli labels: risk of a single tree)**.** *Let K be a compact in $\mathbb{R}^d, d \in \mathbb{N}$. Let $X, X_1, ..., X_n, n \in N^*$ be i.i.d. random variables uniformly distributed over K, $Y, Y_1, ..., Y_n$ i.i.d Bernoulli variables of parameter $p \in [0,1]$ which can be considered as the labels of $X, X_1, ..., X_n$. We denote by $r_{0,k,n}, k \in \mathbb{N}^*$ a single tree of depth k. Then we have, for all $k \in N^*$,*

(i)

$$\mathbb{E}\left[(r_{0,0,n}(X) - r(X))^2\right] = \frac{p(1-p)}{n} \tag{5}$$

(ii)

$$2^k \cdot \frac{p(1-p)}{n} + \left(p^2 - \frac{2^k}{n}\right)(1-2^{-k})^n \leq \mathbb{E}\left[(r_{0,k,n}(X) - r(X))^2\right] \leq 2^{k+1}\cdot\frac{p(1-p)}{n} + p^2(1-2^{-k})^n \tag{6}$$

Proof. (i) In the case $k=0$, $r_{0,0,n}$ simply computes the mean of all the (Y_i)'s over K:

$$\mathbb{E}\left[(r_{0,0,n}(X) - r(X))^2\right] = \mathbb{E}\left[\left(\frac{1}{n}\sum_i Y_i - p\right)^2\right] \tag{7}$$
$$= \mathbb{E}\left[\frac{1}{n^2}\sum_i (Y_i - p)^2\right] \qquad (Y_i \text{ independent}) \tag{8}$$
$$= \frac{p(1-p)}{n}. \tag{9}$$

(ii)

$$\mathbb{E}\left[(r_{0,k,n}(X)-r(X))^2\right]=\mathbb{E}\left[\left(\frac{1}{N_n(L_n(X))}\sum_{X_i\in L_n(X)}Y_i-p\right)^2\mathbb{1}_{N_n(L_n(X))>0}\right]+p^2\mathbb{P}\left(N_n(L_n(X))=0\right) \tag{10}$$

$$=\mathbb{E}\left[\frac{\mathbb{1}_{N_n(L_n(X))>0}}{N_n(L_n(X))^2}\sum_{X_i\in L_n(X)}(Y_i-p)^2\right]+p^2\mathbb{P}\left(N_n(L_n(X))=0\right) \tag{11}$$

$$=p(1-p)\mathbb{E}\left[\frac{\mathbb{1}_{N_n(L_n(X))>0}}{N_n(L_n(X))}\right]+p^2(1-2^{-k})^n \tag{12}$$

Noticing that $N_n(L_n(X))$ is a binomial $\mathfrak{B}(n,\frac{1}{2^k})$, we obtain the upper bound using Lemma S5 (i) :

$$\mathbb{E}\left[\frac{\mathbb{1}_{N_n(L_n(X))>0}}{N_n(L_n(X))}\right]\leq 2\cdot\frac{2^k}{n} \tag{13}$$

the lower bound is immediately obtained by applying Lemma S5, (i):

$$\mathbb{E}\left[\frac{\mathbb{1}_{N_n(L_n(X))>0}}{N_n(L_n(X))}\right]\geq\frac{2^k}{n}\left(1-(1-2^{-k})^n\right) \tag{14}$$

□

S3 Proof of Lemma 1

Note that

$$\mathbb{E}[(\hat{r}_{k,1,\infty}(X)-r(X))^2]=\mathbb{E}[(\hat{r}_{k,1,\infty}(X)-r(X))^2\mathbb{1}_{X\in\mathcal{B}}] \tag{15}$$

$$+\mathbb{E}[(\hat{r}_{k,1,\infty}(X)-r(X))^2\mathbb{1}_{X\in\mathcal{W}}]. \tag{16}$$

Now, we analyze the first term in Equation (16). We have

$$\mathbb{E}[(\hat{r}_{k,1,\infty}(X)-r(X))^2\mathbb{1}_{X\in\mathcal{B}}]=\mathbb{E}[\mathbb{E}[(\hat{r}_{k,1,\infty}(X)-r(X))^2\mathbb{1}_{X\in\mathcal{B}}|N_{\mathcal{B}}]] \tag{17}$$

$$=\sum_{i,j}\mathbb{E}[\mathbb{E}[\Big(\frac{1}{N_{\mathcal{B}}}\sum_{i',j'\in\mathcal{B}}p_{i'j'}-p_{ij}\Big)^2\mathbb{1}_{X\in C_{ij}\cap\mathcal{B}}|N_{\mathcal{B}}]]$$

$$=\sum_{i,j}\mathbb{E}\left[\mathbb{E}[\mathbb{1}_{X\in C_{ij}}\mathbb{1}_{X\in\mathcal{B}}\mathbb{1}_{N_{\mathcal{B}}>0}\Big(\frac{1}{N_{\mathcal{B}}}\sum_{i',j'\in\mathcal{B}}p_{i'j'}-p_{ij}\Big)^2|N_{\mathcal{B}}]\right]$$

$$+\sum_{i,j}\mathbb{E}\left[\mathbb{E}\left[\mathbb{1}_{X\in C_{i,j}\cap\mathcal{B}}\mathbb{1}_{N_{\mathcal{B}}=0}p_{i,j}^2\mid N_{\mathcal{B}}\right]\right]. \tag{18}$$

We begin with the second term in Equation (18). We have, for all i,j,

$$\mathbb{E}\left[\mathbb{E}\left[\mathbb{1}_{X\in C_{i,j}\cap\mathcal{B}}\mathbb{1}_{N_{\mathcal{B}}=0}p_{i,j}^2\mid N_{\mathcal{B}}\right]\right]=\mathbb{E}\left[\mathbb{1}_{X\in C_{i,j}}\mathbb{1}_{N_{\mathcal{B}}=0}\mathbb{E}\left[p_{i,j}^2\mathbb{1}_{X\in\mathcal{B}}\mid X,N_{\mathcal{B}}\right]\right] \tag{19}$$

$$=\mathbb{E}\left[\mathbb{1}_{X\in C_{i,j}}\mathbb{1}_{N_{\mathcal{B}}=0}\mathbb{E}\left[p_{i,j}^2\mathbb{1}_{p_{i,j}\geq\frac{1}{2}}\right]\right]. \tag{20}$$

As $p_{i,j}$ is drawn uniformly in $[0,1]$,

$$\mathbb{E}\left[p_{i,j}^2\mathbb{1}_{p_{i,j}\geq\frac{1}{2}}\right]=\mathbb{E}\left[p_{i,j}^2\mid p_{i,j}\geq\frac{1}{2}\right]\mathbb{P}\left(p_{i,j}\geq\frac{1}{2}\right)$$

$$=\frac{7}{24}.$$

Therefore,

$$\mathbb{E}\left[\mathbb{E}\left[\mathbb{1}_{X\in C_{i,j}\cap\mathcal{B}}\mathbb{1}_{N_\mathcal{B}=0}p_{i,j}^2 \mid N_\mathcal{B}\right]\right] = \frac{7}{24}\mathbb{P}\left(X\in C_{i,j}\right)\mathbb{P}\left(N_\mathcal{B}=0\right) \tag{21}$$

$$= \frac{1}{2^{2^{k^\star}}}\frac{7}{24}. \tag{22}$$

Regarding the first term of Equation (18),

$$\mathbb{E}[\mathbb{1}_{X\in C_{ij}}\mathbb{1}_{X\in\mathcal{B}}\mathbb{1}_{N_\mathcal{B}>0}\Big(\frac{1}{N_\mathcal{B}}\sum_{i',j'\in\mathcal{B}}p_{i'j'}-p_{ij}\Big)^2|N_\mathcal{B}]$$

$$= \mathbb{E}\left[\mathbb{1}_{X\in C_{ij}}\mathbb{E}\left[\mathbb{1}_{X\in\mathcal{B}}\mathbb{1}_{N_\mathcal{B}>0}\left(\frac{1}{N_\mathcal{B}}\sum_{i',j'}(p_{i',j'}-p_{i,j})\right)^2 \mid X, N_\mathcal{B}\right]\right] \tag{23}$$

$$= \mathbb{E}\left[\mathbb{1}_{X\in C_{ij}}\right]\mathbb{E}\left[\mathbb{1}_{p_{i,j}\geq\frac{1}{2}}\mathbb{1}_{N_\mathcal{B}>0}\left(\frac{1}{N_\mathcal{B}}\sum_{i',j'\in\mathcal{B}}(p_{i',j'}-p_{i,j})\right)^2 \mid N_\mathcal{B}\right] \tag{24}$$

$$= \mathbb{E}\left[\mathbb{1}_{X\in C_{ij}}\right]\mathbb{E}\left[\frac{1}{N_\mathcal{B}^2}\left(\sum_{\substack{i',j',i'',j''\in\mathcal{B}\\(i',j')\neq(i'',j'')}}(p_{i'j'}-p_{ij})(p_{i''j''}-pij)+\sum_{\substack{i',j'\in\mathcal{B},\\(i',j'\neq(i,j))}}(p_{i'j'}-p_{ij})^2\right) \mid N_\mathcal{B}, N_\mathcal{B}>0, p_{i,j}\geq\frac{1}{2}\right]$$

$$\cdot\,\mathbb{P}\left(p_{i,j}\geq\frac{1}{2}\cap N_\mathcal{B}>0\right). \tag{25}$$

Recall that p_{ij} is drawn uniformly over $[0,1]$. Therefore, $\mathbb{P}\left(p_{i,j}\geq\frac{1}{2}\cap N_\mathcal{B}>0\right)=\mathbb{P}\left(p_{i,j}\geq\frac{1}{2}\right)=\frac{1}{2}$. Thus,

$$\mathbb{E}[\mathbb{1}_{X\in C_{ij}}\mathbb{1}_{X\in\mathcal{B}}\mathbb{1}_{N_\mathcal{B}>0}\Big(\frac{1}{N_\mathcal{B}}\sum_{i',j'\in\mathcal{B}}p_{i'j'}-p_{ij}\Big)^2|N_\mathcal{B}]$$

$$= \frac{1}{2^{k^\star}}\mathbb{E}\left[\frac{1}{N_\mathcal{B}^2}(N_\mathcal{B}-1)(N_\mathcal{B}-2)\mathrm{Var}(p_{i,j}\mid p_{i,j}\geq\frac{1}{2})+\sum_{\substack{i',j'\in\mathcal{B},\\(i',j'\neq(i,j))}}2\mathrm{Var}(p_{i,j}\mid p_{i,j}\geq\frac{1}{2}) \mid N_\mathcal{B}, N_\mathcal{B}>0, p_{i,j}\geq\frac{1}{2}\right]$$

$$= \frac{1}{2^{k^\star}}\mathbb{E}\left[\frac{1}{N_\mathcal{B}^2}\left((N_\mathcal{B}-1)(N_\mathcal{B}-2)\frac{1}{48}+2\frac{1}{48}(N_\mathcal{B}-1)\right) \mid N_\mathcal{B}, N_\mathcal{B}>0\right]\mathbb{P}\left(p_{i,j}\geq\frac{1}{2}\cap N_\mathcal{B}>0\right)$$

$$= \frac{1}{2^{k^\star+1}}\mathbb{E}\left[\frac{1}{48N_\mathcal{B}^2}\left(N_\mathcal{B}^2-N_\mathcal{B}\right) \mid N_\mathcal{B}, N_\mathcal{B}>0\right]$$

$$= \frac{1}{48\cdot 2^{k^\star+1}}\left(1-\mathbb{E}\left[\frac{1}{N_\mathcal{B}} \mid N_\mathcal{B}, N_\mathcal{B}>0\right]\right).$$

We now have:

$$\mathbb{E}\left[\mathbb{E}[\mathbb{1}_{X\in C_{ij}}\mathbb{1}_{X\in\mathcal{B}}\mathbb{1}_{N_\mathcal{B}>0}\Big(\frac{1}{N_\mathcal{B}}\sum_{i',j'\in\mathcal{B}}p_{i'j'}-p_{ij}\Big)^2|N_\mathcal{B}]\right] = \mathbb{E}\left[\frac{1}{48\cdot 2^{k^\star+1}}\left(1-\mathbb{E}\left[\frac{1}{N_\mathcal{B}} \mid N_\mathcal{B}, N_\mathcal{B}>0\right]\right)\right]$$

$$= \frac{1}{48\cdot 2^{k^\star+1}}\left(1-\mathbb{E}\left[\frac{1}{N_\mathcal{B}} \mid N_\mathcal{B}>0\right]\right).$$

Notice that $N_\mathcal{B}$ is a binomial variable of parameters $2^{k^\star}, 1/2$. Thus we can apply Lemma S5 to deduce

$$\mathbb{E}\left[\frac{1}{N_\mathcal{B}} \mid N_\mathcal{B}>0\right] = \mathbb{E}\left[\frac{\mathbb{1}_{Z>0}}{Z}\right]\frac{1}{\mathbb{P}\left(Z>0\right)}$$

$$\leq \frac{4}{2^{k^\star}+1}\frac{1}{\mathbb{P}\left(Z>0\right)}$$

Moreover, as $\mathbb{P}(Z > 0) \geq \frac{1}{2}$, we have:

$$\mathbb{E}\left[\mathbb{E}[\mathbb{1}_{X\in C_{ij}}\mathbb{1}_{X\in\mathcal{B}}\mathbb{1}_{N_\mathcal{B}>0}\Big(\frac{1}{N_\mathcal{B}}\sum_{i',j'\in\mathcal{B}} p_{i'j'} - p_{ij}\Big)^2 | N_\mathcal{B}]\right] \geq \frac{1}{48\cdot 2^{k^\star+1}}\left(1 - \frac{8}{2^{k^\star}+1}\right)$$

In the end, the first term of Equation (16) verifies

$$\mathbb{E}[(\hat{r}_{k,1,\infty}(X) - r(X))^2\mathbb{1}_{X\in\mathcal{B}}] \geq \frac{1}{2}\Big(\frac{1}{48}\Big(1 - \frac{8}{2^{k^\star}-1}\Big)\Big) + \frac{1}{2^{2^{k^\star}}}\frac{7}{24}.$$

Similar computations show that the second term of Equation (26) verifies:

$$\begin{aligned}
\mathbb{E}[(\hat{r}_{k,1,\infty}(X) - r(X))^2\mathbb{1}_{X\in\mathcal{W}}]. &\geq \frac{1}{2}\Big(\frac{1}{48}\Big(1 - \frac{8}{2^{k^\star}-1}\Big)\Big) + \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}_{X\in C_{ij}\cap\mathcal{W}}\mathbb{1}_{N_\mathcal{W}=0}p_{ij}^2 \mid N_\mathcal{W}\right]\right] && (26)\\
&\geq \frac{1}{2}\Big(\frac{1}{48}\Big(1 - \frac{8}{2^{k^\star}-1}\Big)\Big) + \frac{1}{2^{k^\star}}\frac{1}{2^{2^{k^\star}}}\mathbb{E}\left[p_{ij}^2 \mid p_{ij} < \frac{1}{2}\right] && (27)\\
&\geq \frac{1}{2}\Big(\frac{1}{48}\Big(1 - \frac{8}{2^{k^\star}-1}\Big)\Big) + \frac{1}{2^{2^{k^\star}}}\frac{1}{12}. && (28)
\end{aligned}$$

All in all, we have

$$\mathbb{E}[(\hat{r}_{k,1,\infty}(X) - r(X))^2] \geq \frac{1}{48}\Big(1 - \frac{8}{2^{k^\star}-1}\Big) + \frac{1}{2^{2^{k^\star}}}\frac{9}{24}.$$

S4 Proof of Lemma 2

First, note that since we are in an infinite sample regime, the risk of our estimators is equal to their bias term. We can thus work with the true distribution instead of a finite data set.

(i) The risk of a second-layer tree cutting k' times, $k' \geq k^\star$ along the raw features equals 0 (thus being minimal) as each leaf is included in a cell. We now exhibit one configuration for which any second-layer tree of depth $k' < k^\star$ is biased. We consider the balanced chessboard with parameters $k^\star$, $N_\mathcal{B} = 2^{k^\star-1}$ and p, defined in Proposition 3 and shown in Figure 7. For all $k < k^\star$, each leaf of the first tree contains exactly half black and half white cells, thus predicting $1/2$ and having a risk of $(p - \frac{1}{2})^2$. Therefore a second-layer tree building on raw features only would predict $1/2$ everywhere and would also be biased. If the second-layer tree performs a cut on the new feature provided by the first-layer tree, it creates two leaves: all the leaves where the prediction of the first tree is greater than or equal to $1/2$ are gathered in the right leaf, all the other leaves are gathered in the left leaf. The left leaf is empty and the prediction of the second-layer tree is also $1/2$ everywhere. Any new cut along the new feature would create one leaf predicting $1/2$ on $[0,1]^2$ and other leaves being empty. In any case, the second-layer tree is biased. Thus the minimal risk for all configurations is obtained by a second-layer tree of depth $k' \geq k^\star$ which cuts along the raw features only.

(ii) When $k \geq k^\star$, the first tree is unbiased since each of its leaves is included in only one chessboard data cell. Splitting on the new feature in the second-layer tree induces a separation between cells for which $\mathbb{P}[Y = 1|X \in C] = p$ and cells for which $\mathbb{P}[Y = 1|X \in C] = 1 - p$ since $p \neq 1/2$. Taking the expectation of Y on these two regions leads to a shallow tree network of risk zero.

S5 Proof of Proposition 3

S5.1 Proof of statement 1.: risk of a single tree

As in the precedent proof, we distinguish the case where the cell containing X might be empty, in such a case the tree will predict 0:

$$\begin{aligned}
&\mathbb{E}\left[(r_{k,0,n}(X) - r(X))^2)\right] \\
&\quad = \mathbb{E}\left[(r_{k,0,n}(X) - r(X))^2 \mathbb{1}_{N_n(L_n(X))>0}\right] + \mathbb{E}\left[(r(X))^2 \mathbb{1}_{N_n(L_n(X))=0}\right] &(29)\\
&\quad = \mathbb{E}\left[(r_{k,0,n}(X) - r(X))^2 \mathbb{1}_{N_n(L_n(X))>0}\right] + (p^2 + (1-p)^2)\frac{(1-2^{-k})^n}{2}. &(30)
\end{aligned}$$

We denote by $L_1, ..., L_{2^k}$ the leaves of the tree. Let $b \in \{1, \dots, 2^k\}$ such that L_b belongs to $\mathcal{B}$. We have

$$\begin{aligned}
&\mathbb{E}\left[(r_{k,0,n}(X) - p)^2) \mathbb{1}_{X\in\mathcal{B}} \mathbb{1}_{N_n(L_n(X))>0}\right] \\
&\quad = \sum_{L_j \subset \mathcal{B}} \mathbb{E}\left[\left(\frac{\mathbb{1}_{N_n(L_j)>0}}{N_n(L_j)} \sum_{X_i \in L_j} (Y_i - p)\right)^2 \mathbb{1}_{X\in L_j}\right] &(31)\\
&\quad = \frac{2^k}{2} \cdot \mathbb{E}\left[\left(\frac{\mathbb{1}_{N_n(L_b)>0}}{N_n(L_b)} \sum_{X_i \in L_b} (Y_i - p)\right)^2\right] \mathbb{P}\left(X \in L_b\right) &(32)\\
&\quad = \frac{1}{2}\mathbb{E}\left[\left(\frac{\mathbb{1}_{N_n(L_b)>0}}{N_n(L_b)} \sum_{X_i \in L_b} (Y_i - p)\right)^2\right] &(33)\\
&\quad = \frac{1}{2}\mathbb{E}\left[\frac{\mathbb{1}_{N_n(L_b)>0}}{N_n(L_b)^2}\mathbb{E}\left[\left(\sum_{X_i \in L_b} (Y_i - p)\right)^2 | N_n(L_b)\right]\right] &(34)\\
&\quad = \frac{1}{2}\mathbb{E}\left[\frac{\mathbb{1}_{N_n(L_b)>0}}{N_n(L_b)^2}\mathbb{E}\left[\sum_{X_i \in L_b} (Y_i - p)^2 | N_n(L_b)\right]\right] \qquad \text{(by independence of the } Y_i\text{)} &(35)\\
&\quad = \frac{1}{2}\mathbb{E}\left[\frac{\mathbb{1}_{N_n(L_b)>0}}{N_n(L_b)} p(1-p)\right]. &(36)
\end{aligned}$$

Remark that the above computation holds when $X \in \mathcal{W}$ after replacing p by $(1-p)$, $\mathcal{B}$ by $\mathcal{W}$ and L_b by L_w: indeed when Y is a Bernoulli random variable, Y and $1-Y$ have the same variance. Hence, using Equation (30), the computation in (36) and its equivalence for $\mathcal{W}$, we obtain

$$\begin{aligned}
&\mathbb{E}\left[(r_{k,0,n}(X) - r(X))^2)\right] \\
&= \frac{1}{2}\mathbb{E}\left[\frac{\mathbb{1}_{N_n(L_b)>0}}{N_n(L_b)} p(1-p)\right] + \frac{1}{2}\mathbb{E}\left[\frac{\mathbb{1}_{N_n(L_w)>0}}{N_n(L_w)} p(1-p)\right] + (p^2 + (1-p)^2)\frac{(1-2^{-k})^n}{2} \\
&= p(1-p)\mathbb{E}\left[\frac{\mathbb{1}_{N_n(L_w)>0}}{N_n(L_w)}\right] + (p^2 + (1-p)^2)\frac{(1-2^{-k})^n}{2},
\end{aligned}$$

since $N_n(L_b)$ and $N_n(L_w)$ are both binomial random variables $\mathfrak{B}(n, \frac{1}{2^k})$. Therefore we can conclude using Lemma S5 (i):

$$\mathbb{E}\left[(r_{k,0,n}(X) - r(X))^2)\right] \leq \frac{2^k p(1-p)}{n+1} + \left(p^2 + (1-p)^2\right)\frac{(1-2^{-k})^n}{2}$$

and

$$\mathbb{E}\left[(r_{k,0,n}(X) - r(X))^2)\right] \geq \frac{2^{k-1} p(1-p)}{n+1} + \left(p^2 + (1-p)^2 - \frac{2^k p(1-p)}{n+1}\right)\frac{(1-2^{-k})^n}{2}.$$

S5.2 Proof of statement 2.: risk of a shallow tree network

Let $k \in \mathbb{N}$. Denote by $\mathcal{L}_k = \{L_i, i = 1, \dots, 2^k\}$ the set of all leaves of the encoding tree (of depth k). We let $\mathcal{L}_{\tilde{\mathcal{B}}_k}$ be the set of all cells of the encoding tree containing at least one observation, and such that the empirical probability of Y being equal to one in the cell is larger than $1/2$, i.e.

$$\tilde{\mathcal{B}}_k = \cup_{L \in \mathcal{L}_{\tilde{\mathcal{B}}_k}} \{x, x \in L\}$$

$$\mathcal{L}_{\tilde{\mathcal{B}}_k} = \{L \in \mathcal{L}_k, N_n(L) > 0, \frac{1}{N_n(L)} \sum_{X_i \in L} Y_i \geq \frac{1}{2}\}.$$

Accordingly, we let the part of the input space corresponding to $\mathcal{L}_{\tilde{\mathcal{B}}_k}$ as

$$\tilde{\mathcal{B}}_k = \cup_{L \in \mathcal{L}_{\tilde{\mathcal{B}}_k}} \{x, x \in L\}$$

Similarly,

$$\mathcal{L}_{\tilde{\mathcal{W}}_k} = \{L \in \mathcal{L}_k, N_n(L) > 0, \frac{1}{N_n(L)} \sum_{X_i \in L} Y_i < \frac{1}{2}\}.$$

and

$$\tilde{\mathcal{W}}_k = \cup_{L \in \mathcal{L}_{\tilde{\mathcal{W}}_k}} \{x, x \in L\}$$

S5.2.1 Proof of 2. (upper-bound)

Recall that $k \geq k^\star$. In this case, each leaf of the encoding tree is included in a chessboard cell. As usual,

$$\mathbb{E}\left[(r_{k,1,n}(X) - r(X))^2)\right] = \mathbb{E}\left[(r_{k,1,n}(X) - r(X))^2 \mathbb{1}_{N_n(L_n(X))>0}\right] + \frac{p^2 + (1-p)^2}{2}\left(1 - \frac{1}{2^k}\right)^n. \quad (37)$$

Note that

$$\begin{aligned}
&\mathbb{E}\left[(r_{k,1,n}(X) - r(X))^2 \mathbb{1}_{N_n(L_n(X))>0}\right] \\
&= \mathbb{E}\left[\left(\frac{1}{N_n(\tilde{\mathcal{B}}_k)} \sum_{X_i \in \tilde{\mathcal{B}}_k} Y_i - p\right)^2 \mathbb{1}_{X \in \mathcal{B}} \mathbb{1}_{X \in \tilde{\mathcal{B}}_k}\right] + \mathbb{E}\left[\left(\frac{1}{N_n(\tilde{\mathcal{W}}_k)} \sum_{X_i \in \tilde{\mathcal{W}}_k} Y_i - p\right)^2 \mathbb{1}_{X \in \mathcal{B}} \mathbb{1}_{X \in \tilde{\mathcal{W}}_k}\right] \\
&+ \mathbb{E}\left[\left(\frac{1}{N_n(\tilde{\mathcal{B}}_k)} \sum_{X_i \in \tilde{\mathcal{B}}_k} Y_i - (1-p)\right)^2 \mathbb{1}_{X \in \mathcal{W}} \mathbb{1}_{X \in \tilde{\mathcal{B}}_k}\right] + \mathbb{E}\left[\left(\frac{1}{N_n(\tilde{\mathcal{W}}_k)} \sum_{X_i \in \tilde{\mathcal{W}}_k} Y_i - (1-p)\right)^2 \mathbb{1}_{X \in \mathcal{W}} \mathbb{1}_{X \in \tilde{\mathcal{W}}_k}\right] \\
&\leq \frac{1}{2}\mathbb{E}\left[\left(\frac{1}{N_n(\tilde{\mathcal{B}}_k)} \sum_{X_i \in \tilde{\mathcal{B}}_k} Y_i - p\right)^2 \mathbb{1}_{N_n(\tilde{\mathcal{B}}_k)>0}\right] + \frac{1}{2}\mathbb{E}\left[\left(\frac{1}{N_n(\tilde{\mathcal{W}}_k)} \sum_{X_i \in \tilde{\mathcal{W}}_k} Y_i - (1-p)\right)^2 \mathbb{1}_{N_n(\tilde{\mathcal{W}}_k)>0}\right] \\
&\qquad + \mathbb{E}\left[\mathbb{1}_{X \in \mathcal{B}, X \in \tilde{\mathcal{W}}_k}\right] + \mathbb{E}\left[\mathbb{1}_{X \in \mathcal{W}, X \in \tilde{\mathcal{B}}_k}\right]. \quad (38)
\end{aligned}$$

Let L be a generic cell. The third term in (38) can be upper-bounded as follows:

$$\begin{aligned}
\mathbb{E}\left[\mathbb{1}_{X \in \mathcal{B}, X \in \tilde{\mathcal{W}}_k}\right] &= \sum_{j=1}^{2^k} \mathbb{E}\left[\mathbb{1}_{X \in L_j} \mathbb{1}_{L_j \subset \tilde{\mathcal{W}}_k \cap \mathcal{B}}\right] && (39) \\
&= \sum_{j=1}^{2^k} \mathbb{P}\left(X \in L_j\right) \mathbb{P}\left(L_j \subset \tilde{\mathcal{W}}_k \cap \mathcal{B}\right) && (40) \\
&= \sum_{j=1}^{2^k} \mathbb{P}\left(X \in L_j\right) \mathbb{P}\left(L_j \subset \tilde{\mathcal{W}}_k \mid L_j \subset \mathcal{B}\right) \mathbb{P}\left(L_j \subset \mathcal{B}\right) && (41) \\
&= \frac{1}{2}\mathbb{P}\left(L \subset \tilde{\mathcal{W}}_k \mid L \subset \mathcal{B}\right), && (42)
\end{aligned}$$

by symmetry. Now,

$$\mathbb{P}\left(L \subset \tilde{\mathcal{W}}_k \mid L \subset \mathcal{B}\right) = \mathbb{P}\left(\frac{1}{N_n(L)} \sum_{X_i \in L} \mathbb{1}_{Y_i=0} > \frac{1}{2} \mid L \subset \mathcal{B}\right) \tag{43}$$

$$\leq \mathbb{E}\left[\mathbb{P}\left(\frac{1}{N_n(L)} \sum_{X_i \in L, L \subset \mathcal{B}} \mathbb{1}_{Y_i=0} - (1-p) \geq \frac{1}{2} - (1-p) | N_n(L), L \subset \mathcal{B}\right) \mid L \subset \mathcal{B}\right] \tag{44}$$

$$\leq \mathbb{E}\left[e^{-2N_n(L)(p-\frac{1}{2})^2}\right] \tag{45}$$

(according to Hoeffding's inequality)

$$= \prod_{i=1}^{n} \mathbb{E}\left[e^{-2(p-\frac{1}{2})^2 \mathbb{1}_{X_i \in L}}\right] \tag{46}$$

(by independence of X_i's)

$$= \left(\frac{e^{-2(p-\frac{1}{2})^2}}{2^k} + 1 - \frac{1}{2^k}\right)^n. \tag{47}$$

Consequently,

$$\mathbb{E}\left[\mathbb{1}_{X \in \mathcal{B}, X \in \tilde{\mathcal{W}}_k}\right] \leq \frac{1}{2}\left(\frac{e^{-2(p-\frac{1}{2})^2}}{2^k} + 1 - \frac{1}{2^k}\right)^n.$$

Similar calculations show that

$$\begin{aligned}\mathbb{E}\left[\mathbb{1}_{X \in \mathcal{W}, X \in \tilde{\mathcal{B}}_k}\right] &= \frac{1}{2}\mathbb{P}\left(L \subset \tilde{\mathcal{B}}_k \mid L \subset \mathcal{W}\right) \\ &\leq \frac{1}{2}\left(\frac{e^{-2(p-\frac{1}{2})^2}}{2^k} + 1 - \frac{1}{2^k}\right)^n. \end{aligned} \tag{48}$$

Therefore,

$$\begin{aligned}&\mathbb{E}\left[(r_{k,1,n}(X) - r(X))^2)\right] \\ &\leq \frac{1}{2}\mathbb{E}\left[\left(\frac{1}{N_n(\tilde{\mathcal{B}}_k)} \sum_{X_i \in \tilde{\mathcal{B}}_k} Y_i - p\right)^2 \mathbb{1}_{N_n(\tilde{\mathcal{B}}_k)>0}\right] + \frac{1}{2}\mathbb{E}\left[\left(\frac{1}{N_n(\tilde{\mathcal{W}}_k)} \sum_{X_i \in \tilde{\mathcal{W}}_k} Y_i - (1-p)\right)^2 \mathbb{1}_{N_n(\tilde{\mathcal{W}}_k)>0}\right] \\ &\quad + \left(\frac{e^{-2(p-\frac{1}{2})^2}}{2^k} + 1 - \frac{1}{2^k}\right)^n + \frac{p^2 + (1-p)^2}{2}\left(1 - \frac{1}{2^k}\right)^n. \end{aligned} \tag{49}$$

Now, the first term in (49) can be written as

$$\mathbb{E}\left[\left(\frac{1}{N_n(\tilde{\mathcal{B}}_k)} \sum_{X_i \in \tilde{\mathcal{B}}_k} Y_i - p\right)^2 \mathbb{1}_{N_n(\tilde{\mathcal{B}}_k)>0}\right] \tag{50}$$

$$= \mathbb{E}\left[\left(\frac{1}{N_n(\tilde{\mathcal{B}}_k)} \sum_{X_i \in \tilde{\mathcal{B}}_k} Y_i - p\right)^2 \mathbb{1}_{N_n(\tilde{\mathcal{B}}_k)>0} \mathbb{1}_{\mathcal{B}=\tilde{\mathcal{B}}_k}\right] + \mathbb{E}\left[\left(\frac{1}{N_n(\tilde{\mathcal{B}}_k)} \sum_{X_i \in \tilde{\mathcal{B}}_k} Y_i - p\right)^2 \mathbb{1}_{N_n(\tilde{\mathcal{B}}_k)>0} \mathbb{1}_{\mathcal{B} \neq \tilde{\mathcal{B}}_k}\right] \tag{51}$$

$$\leq \mathbb{E}\left[\left(\frac{1}{N_n(\tilde{\mathcal{B}}_k)} \sum_{X_i \in \tilde{\mathcal{B}}_k} Y_i - p\right)^2 \mathbb{1}_{N_n(\tilde{\mathcal{B}}_k)>0} \mathbb{1}_{\mathcal{B}=\tilde{\mathcal{B}}_k}\right] + \mathbb{P}\left(\mathcal{B} \neq \tilde{\mathcal{B}}_k\right) \tag{52}$$

Now, using a union bound, we obtain

$$\mathbb{P}\left(\mathcal{B} \neq \tilde{\mathcal{B}}_k\right) \leq \sum_{L_j \subset \mathcal{B}} \mathbb{P}\left(L_j \not\subset \tilde{\mathcal{B}}_k\right) + \sum_{L_j \subset \mathcal{W}} \mathbb{P}\left(L_j \subset \tilde{\mathcal{B}}_k\right) \tag{53}$$

$$\leq \frac{2^k}{2} \cdot \mathbb{P}\left(L \not\subset \tilde{\mathcal{B}}_k \mid L \subset \mathcal{B}\right) + \frac{2^k}{2} \cdot \mathbb{P}\left(L \subset \tilde{\mathcal{B}}_k \mid L \subset \mathcal{W}\right) \tag{54}$$

$$\leq 2^k \left(\frac{e^{-2(p-\frac{1}{2})^2}}{2^k} + 1 - \frac{1}{2^k}\right)^n, \tag{55}$$

according to (47) and (48). Additionally, the left term in (52) satisfies

$$\mathbb{E}\left[\left(\frac{1}{N_n(\tilde{\mathcal{B}}_k)} \sum_{X_i \in \tilde{\mathcal{B}}_k} Y_i - p\right)^2 \mathbb{1}_{N_n(\tilde{\mathcal{B}}_k)>0} \mathbb{1}_{\mathcal{B}=\tilde{\mathcal{B}}_k}\right] \leq \mathbb{E}\left[\left(\frac{1}{N_n(\mathcal{B})} \sum_{X_i \in \mathcal{B}} Y_i - p\right)^2 \mathbb{1}_{N_n(\mathcal{B})>0}\right] \tag{56}$$

$$\leq \mathbb{E}\left[\frac{\mathbb{1}_{N_n(\mathcal{B})>0}}{N_n(\mathcal{B})^2}\left(\sum_{X_i \in \mathcal{B}} Y_i - pN_n(\mathcal{B})\right)^2\right] \tag{57}$$

$$= p(1-p)\mathbb{E}\left[\frac{\mathbb{1}_{N_n(\mathcal{B})>0}}{N_n(\mathcal{B})}\right], \tag{58}$$

noticing that the square term of (57) is nothing but the conditional variance of a binomial distribution $B(N_n(\mathcal{B}), p)$. By Lemma S5 (i) on $N_n(\mathcal{B})$ which is a binomial random variable $B(n, p)$ with $p = 1/2$ (exactly half of the cells are black),

$$\mathbb{E}\left[\left(\frac{1}{N_n(\tilde{\mathcal{B}}_k)} \sum_{X_i \in \tilde{\mathcal{B}}_k} Y_i - p\right)^2 \mathbb{1}_{N_n(\tilde{\mathcal{B}}_k)>0} \mathbb{1}_{N_n(\tilde{\mathcal{B}}_k)>0}\right] \leq \frac{2p(1-p)}{n+1}.$$

Hence

$$\mathbb{E}\left[\left(\frac{1}{N_n(\tilde{\mathcal{B}}_k)} \sum_{X_i \in \tilde{\mathcal{B}}_k} Y_i - p\right)^2 \mathbb{1}_{\mathcal{B}=\tilde{\mathcal{B}}_k}\right] \leq \frac{2p(1-p)}{n+1} + 2^k \left(\frac{e^{-2(p-\frac{1}{2})^2}}{2^k} + 1 - \frac{1}{2^k}\right)^n. \tag{59}$$

Similarly,

$$\mathbb{E}\left[\left(\frac{1}{N_n(\tilde{\mathcal{W}}_k)} \sum_{X_i \in \tilde{\mathcal{W}}_k} Y_i - (1-p)\right)^2 \mathbb{1}_{N_n(\tilde{\mathcal{W}}_k)>0}\right] \leq \frac{2p(1-p)}{n+1} + 2^k \left(\frac{e^{-2(p-\frac{1}{2})^2}}{2^k} + 1 - \frac{1}{2^k}\right)^n. \tag{60}$$

Finally,

Injecting (59) and (60) into (49), we finally get

$$\mathbb{E}\left[(r_{k,1,n}(X) - r(X))^2)\right] \leq \frac{p^2 + (1-p)^2}{2}\left(1 - \frac{1}{2^k}\right)^n + 2^k \cdot \left(\frac{e^{-2(p-\frac{1}{2})^2}}{2^k} + 1 - \frac{1}{2^k}\right)^n + \frac{2p(1-p)}{n+1} + \left(\frac{e^{-2(p-\frac{1}{2})^2}}{2^k} + 1 - \frac{1}{2^k}\right)^n,$$

which concludes this part of the proof.

S5.2.2 Proof of 2. (lower bound)

We have

$$\mathbb{E}\left[(r_{k,1,n}(X) - r(X))^2)\right] = \mathbb{E}\left[(r_{k,1,n}(X) - r(X))^2 \mathbb{1}_{N_n(L_n(X))>0}\right] + \left(\frac{p^2 + (1-p)^2}{2}\right)\left(1 - \frac{1}{2^k}\right)^n,$$

where

$$
\begin{aligned}
&\mathbb{E}\left[(r_{k,1,n}(X)-r(X))^2\mathbb{1}_{N_n(L_n(X))>0}\right] \\
&\geq \mathbb{E}\left[\left(\frac{1}{N_n(\tilde{\mathcal{B}}_k)}\sum_{X_i\in\tilde{\mathcal{B}}_k}Y_i-p\right)^2\mathbb{1}_{X\in\mathcal{B}}\mathbb{1}_{X\in\tilde{\mathcal{B}}_k}\mathbb{1}_{N_n(\tilde{\mathcal{B}}_k)>0}\mathbb{1}_{\mathcal{B}=\tilde{\mathcal{B}}_k}\right] \\
&\quad+\mathbb{E}\left[\left(\frac{1}{N_n(\tilde{\mathcal{W}}_k)}\sum_{X_i\in\tilde{\mathcal{W}}_k}Y_i-(1-p)\right)^2\mathbb{1}_{X\in\mathcal{W}}\mathbb{1}_{X\in\tilde{\mathcal{W}}_k}\mathbb{1}_{N_n(\tilde{\mathcal{W}}_k)>0}\mathbb{1}_{\mathcal{W}=\tilde{\mathcal{W}}_k}\right] \\
&\geq \mathbb{P}\left(X\in\mathcal{B}\right)\mathbb{E}\left[\left(\frac{1}{N_n(\tilde{\mathcal{B}}_k)}\sum_{X_i\in\tilde{\mathcal{B}}_k}Y_i-p\right)^2\mathbb{1}_{\mathcal{B}=\tilde{\mathcal{B}}_k}\mathbb{1}_{N_n(\tilde{\mathcal{B}}_k)>0}\right] \\
&\quad+\mathbb{P}\left(X\in\mathcal{W}\right)\mathbb{E}\left[\left(\frac{1}{N_n(\tilde{\mathcal{W}}_k)}\sum_{X_i\in\tilde{\mathcal{W}}_k}Y_i-(1-p)\right)^2\mathbb{1}_{\mathcal{W}=\tilde{\mathcal{W}}_k}\mathbb{1}_{N_n(\tilde{\mathcal{W}}_k)>0}\right]. \qquad (61)
\end{aligned}
$$

The first expectation term line (61) can be written as

$$
\mathbb{E}\left[\left(\frac{1}{N_n(\tilde{\mathcal{B}}_k)}\sum_{X_i\in\tilde{\mathcal{B}}_k}Y_i-p\right)^2\mathbb{1}_{\mathcal{B}=\tilde{\mathcal{B}}_k}\mathbb{1}_{N_n(\tilde{\mathcal{B}}_k)>0}\right]=\mathbb{P}\left(\mathcal{B}=\tilde{\mathcal{B}}_k\right)\mathbb{E}\left[\left(\frac{1}{N_n(\mathcal{B})}\sum_{X_i\in\mathcal{B}}Y_i-p\right)^2|\mathcal{B}=\tilde{\mathcal{B}}_k\right] \qquad (62)
$$

According to (55),

$$
\mathbb{P}\left(\mathcal{B}=\tilde{\mathcal{B}}_k\right)\geq 1-2^k\cdot\left(1+\frac{e^{-2(p-\frac{1}{2})^2}-1}{2^k}\right)^n. \qquad (63)
$$

Similarly,

$$
\mathbb{P}\left(\mathcal{W}=\tilde{\mathcal{W}}_k\right)\geq 1-2^k\cdot\left(1+\frac{e^{-2(p-\frac{1}{2})^2}-1}{2^k}\right)^n.
$$

Furthermore,

$$
\mathbb{E}\left[\left(\frac{1}{N_n(\mathcal{B})}\sum_{X_i\in\mathcal{B}}Y_i-p\right)^2|\mathcal{B}=\tilde{\mathcal{B}}_k\right]=\mathbb{E}\left[\frac{1}{N_n(\mathcal{B})^2}\mathbb{E}\left[\left(\sum_{X_i\in\mathcal{B}}Y_i-N_n(\mathcal{B})p\right)^2|N_1,...N_{2^k},\mathcal{B}=\tilde{\mathcal{B}}_k\right]|\mathcal{B}=\tilde{\mathcal{B}}_k\right] \qquad (64)
$$

where we let $Z = \sum_{X_i \in \mathcal{B}} Y_i$. A typical bias-variance decomposition yields

$$\mathbb{E}\left[\left(\sum_{X_i \in \mathcal{B}} Y_i - N_n(\mathcal{B})p\right)^2 | N_1, ...N_{2^k}, \mathcal{B} = \tilde{\mathcal{B}}_k\right] \tag{65}$$

$$= \mathbb{E}\left[\left(Z - \mathbb{E}\left[Z \mid N_1, ...N_{2^k}, \tilde{\mathcal{B}}_k = \mathcal{B}\right]\right)^2 + \left(\mathbb{E}\left[Z \mid N_1, ...N_{2^k}, \tilde{\mathcal{B}}_k = \mathcal{B}\right] - N_n(\mathcal{B})p\right)^2 \mid N_1, ...N_{2^k}, \tilde{\mathcal{B}}_k = \mathcal{B}\right] \tag{66}$$

$$\geq \mathbb{E}\left[\left(Z - \mathbb{E}\left[Z \mid N_1, ...N_{2^k}, \tilde{\mathcal{B}}_k = \mathcal{B}\right]\right)^2 \mid N_1, ...N_{2^k}, \tilde{\mathcal{B}}_k = \mathcal{B}\right] \tag{67}$$

$$= \mathbb{E}\left[\left(\sum_{L_j \subset \mathcal{B}} Z_j - \mathbb{E}\left[Z_j \mid N_j, L_j \subset \tilde{\mathcal{B}}_k\right]\right)^2 \mid N_1, ...N_{2^k}, \tilde{\mathcal{B}}_k = \mathcal{B}\right] \tag{68}$$

$$= \sum_{L_j \subset \mathcal{B}} \mathbb{E}\left[\left(Z_j - \mathbb{E}\left[Z_j \mid N_j, L_j \subset \tilde{\mathcal{B}}_k\right]\right)^2 \mid N_j, L_j \subset \tilde{\mathcal{B}}_k\right]$$
$$+ 2 \sum_{L_i, L_j \subset \mathcal{B}, L_i \neq L_j} \mathbb{E}\left[\left(Z_i - \mathbb{E}\left[Z_i \mid N_i, L_i \subset \tilde{\mathcal{B}}_k\right]\right)\left(Z_j - \mathbb{E}\left[Z_j \mid N_j, L_j \subset \tilde{\mathcal{B}}_k\right]\right) \mid N_1, ...N_{2^k}, \tilde{\mathcal{B}}_k = \mathcal{B}\right] \tag{69}$$

$$= \sum_{L_j \subset \mathcal{B}} \mathbb{E}\left[\left(Z_j - \mathbb{E}\left[Z_j \mid N_j, L_j \subset \tilde{\mathcal{B}}_k\right]\right)^2 \mid N_j, L_j \subset \tilde{\mathcal{B}}_k\right]. \tag{70}$$

with $Z_j = \sum_{X_i \in L_j} Y_i$, and $L_1, \ldots, L_{2^k}$ the leaves of the first layer tree. Note that $Z_j | N_j, L_j \subset \mathcal{B}$ are i.i.d binomial variable $\mathfrak{B}(N_j, p)$. In (68) and (69), we used that that given a single leaf $L_j \subset \mathcal{B}$, $\mathbb{E}\left[Z_j \mid N_1, ...N_{2^k}, \tilde{\mathcal{B}}_k = \mathcal{B}\right] = \mathbb{E}\left[Z_j \mid N_j, L_j \subset \tilde{\mathcal{B}}_k\right]$. To obtain (70), we used that conditional to $N_1, ...N_{2^k}, \tilde{\mathcal{B}}_k = \mathcal{B}$, Z_i and Z_j are independent. Therefore the double sum equals 0. Let j be an integer in $\{1, ..., 2^k\}$,

$$\mathbb{E}\left[\left(Z_j - \mathbb{E}\left[Z_j \mid N_j, L_j \subset \tilde{\mathcal{B}}_k\right]\right)^2 \mid N_j, L_j \subset \tilde{\mathcal{B}}_k\right] \tag{71}$$

$$= \mathbb{E}\left[Z_j^2 \mid N_j, L_j \subset \tilde{\mathcal{B}}_k\right] - \mathbb{E}\left[Z_j \mid N_j, L_j \subset \tilde{\mathcal{B}}_k\right]^2 \tag{72}$$

$$\geq \mathbb{E}\left[Z_j^2 \mid N_j\right] - \mathbb{E}\left[Z_j \mid N_j, L_j \subset \tilde{\mathcal{B}}_k\right]^2 \tag{73}$$

$$= N_j p(1-p) + N_j^2 p^2 - \left(N_j p + \frac{N_j}{2}(1-p)\frac{\mathbb{P}\left(Z_j = \frac{N_j}{2} \mid N_j\right)}{\sum_{i=\frac{N_j}{2}}^{N_j} \mathbb{P}\left(Z_j = i\right)}\right)^2 \tag{74}$$

$$\geq N_j(1-p)\left(p - N_j(1-p)\mathbb{P}\left(Z_j = \frac{N_j}{2} \mid N_j\right)^2 - 2N_j p \cdot \mathbb{P}\left(Z_j = \frac{N_j}{2} \mid N_j\right)\right) \tag{75}$$

$$\geq N_j(1-p)\left(p - \frac{N_j(1-p)}{\pi\left(\frac{N_j}{2} + \frac{1}{4}\right)}\left(4p(1-p)\right)^{N_j} - \frac{2N_j}{\sqrt{\pi\left(\frac{N_j}{2} + \frac{1}{4}\right)}}\left(4p(1-p)\right)^{N_j/2}\right) \tag{76}$$

$$\geq N_j p(1-p) - \left(\frac{2(1-p)^2}{\pi} + 2\sqrt{2}(1-p)\right) \cdot N_j^{3/2} \cdot \left(4p(1-p)\right)^{N_j/2}. \tag{77}$$

We deduced Line (73) from the fact that Z_j^2 is a positive random variable, (74) from Lemma (S5) (v), Line (75) from the fact that $p > 1/2$ and Line (76) from the inequality (3) on the binomial coefficient.

Injecting (69) and (77) into (64) yields

$$\mathbb{E}\left[\left(\frac{1}{N_n(\mathcal{B})}\sum_{X_i\in\mathcal{B}}Y_i-p\right)^2|\mathcal{B}=\tilde{\mathcal{B}}_k\right]$$

$$\geq \mathbb{E}\left[\frac{1}{N_n(\mathcal{B}_k)^2}\sum_{L_j\subset\mathcal{B}}\left(N_jp(1-p)-\left(\frac{2(1-p)^2}{\pi}+2\sqrt{2}(1-p)\right)\cdot N_j^{3/2}\cdot(4p(1-p))^{N_j/2}\right)|\mathcal{B}=\tilde{\mathcal{B}}_k\right] \tag{78}$$

$$\geq \mathbb{E}\left[\frac{p(1-p)}{N_n(\mathcal{B})}\mid\mathcal{B}=\tilde{\mathcal{B}}_k\right]-\left(\frac{2(1-p)^2}{\pi}+2\right)\sum_{L_j\subset\mathcal{B}}\mathbb{E}\left[(4p(1-p))^{N_j/2}\mid\mathcal{B}=\tilde{\mathcal{B}}_k\right] \tag{79}$$

$$\geq p(1-p)\mathbb{E}\left[\frac{1}{N_n(\mathcal{B})}\mid\mathcal{B}=\tilde{\mathcal{B}}_k\right]-3\cdot 2^{k-1}\mathbb{E}\left[(4p(1-p))^{N_b/2}\mid\mathcal{B}=\tilde{\mathcal{B}}_k\right] \tag{80}$$

where the last inequality relies on the fact that the $N_j, L_j\subset\mathcal{B}$ are i.i.d, with $b\in 1,...,2^k$ be the index of a cell included in $\mathcal{B}$. N_j is a binomial random variable $\mathfrak{B}(n,2^{-k})$.

$$\mathbb{E}\left[(4p(1-p))^{N_j/2}\mid\mathcal{B}=\tilde{\mathcal{B}}_k\right]\leq\mathbb{E}\left[(4p(1-p))^{N_j/2}\right]\frac{1}{\mathbb{P}\left(\mathcal{B}=\tilde{\mathcal{B}}_k\right)} \tag{81}$$

$$=\left(\sqrt{4p(1-p)}\cdot 2^{-k}+(1-2^{-k})\right)^n\frac{1}{\mathbb{P}\left(\mathcal{B}=\tilde{\mathcal{B}}_k\right)}. \tag{82}$$

From the inequality Line (63), we deduce that as soon as $n\geq\frac{(k+1)\log(2)}{\log(2^k)-\log(e^{-2(p-1/2)^2}-1+2^k)}$,

$$\frac{1}{\mathbb{P}\left(\mathcal{B}=\tilde{\mathcal{B}}_k\right)}\leq 2. \tag{83}$$

Therefore,

$$\mathbb{E}\left[(4p(1-p))^{N_j/2}\mid\mathcal{B}=\tilde{\mathcal{B}}_k\right]\leq 2\left(\sqrt{4p(1-p)}\cdot 2^{-k}+(1-2^{-k})\right)^n. \tag{84}$$

Moreover,

$$\mathbb{E}\left[\frac{1}{N_n(\mathcal{B})}|\mathcal{B}=\tilde{\mathcal{B}}_k\right]\geq\frac{1}{\mathbb{E}\left[N_n(\mathcal{B})|\mathcal{B}=\tilde{\mathcal{B}}_k\right]} \tag{85}$$

$$\geq\frac{\mathbb{P}\left(\mathcal{B}=\tilde{\mathcal{B}}_k\right)}{\mathbb{E}\left[N_n(\mathcal{B})\right]} \tag{86}$$

$$\geq\frac{2}{n}-\frac{2^{k+1}}{n}\left(1+\frac{e^{-2(p-\frac{1}{2})^2}-1}{2^k}\right)^n \tag{87}$$

where the last inequality comes from the probability bound line (63) and the fact that $N_n(\mathcal{B})$ is a binomial random variable $\mathfrak{B}(n,1/2)$.

Finally,

$$\mathbb{E}\left[\left(\frac{1}{N_n(\mathcal{B})}\sum_{X_i\in\mathcal{B}}Y_i-p\right)^2|\mathcal{B}=\tilde{\mathcal{B}}_k\right] \tag{88}$$

$$\geq\frac{2p(1-p)}{n}-3\cdot 2^k\left(1-2^{-k}\left(1-\sqrt{4p(1-p)}\right)\right)^n-\frac{2^{k+1}p(1-p)}{n}\left(1+\frac{e^{-2(p-\frac{1}{2})^2}-1}{2^k}\right)^n. \tag{89}$$

Similarly, regarding the second term of (61), note that $\mathbb{P}\left(\tilde{\mathcal{B}}_k=\mathcal{B}\right)=\mathbb{P}\left(\tilde{\mathcal{W}}_k=\mathcal{W}\right)$ and

$$\mathbb{E}\left[\left(\sum_{X_i\in\mathcal{W}}Y_i-N_n(\mathcal{W})(1-p)\right)^2|N_n(\mathcal{W}),\mathcal{W}=\tilde{\mathcal{W}}_k\right]=\mathbb{E}\left[\left(\sum_{X_i\in\mathcal{W}}\mathbb{1}_{Y_i=0}-N_n(\mathcal{W})p\right)^2|N_n(\mathcal{W}),\mathcal{W}=\tilde{\mathcal{W}}_k\right].$$

Thus we can adapt the above computation to this term :

$$\mathbb{E}\left[\left(\frac{1}{N_n(\mathcal{W})}\sum_{X_i\in\mathcal{W}} Y_i - p\right)^2 |\mathcal{W}=\tilde{\mathcal{W}}_k\right] \tag{90}$$

$$\geq \frac{2p(1-p)}{n} - 3\cdot 2^k\left(1-2^{-k}\left(1-\sqrt{4p(1-p)}\right)\right)^n - \frac{2^{k+1}p(1-p)}{n}\left(1+\frac{e^{-2(p-\frac{1}{2})^2}-1}{2^k}\right)^n. \tag{91}$$

Rearranging all terms proves the result :

$$\begin{aligned}
\mathbb{E}\left[(r_{k,1,n}(X)-r(X))^2\right] &\geq \left(\frac{2p(1-p)}{n} - 2^{k+2}\cdot\left(1-2^{-k}\left(1-\sqrt{4p(1-p)}\right)\right)^n\right.\\
&\quad\left. - \frac{2^{k+1}p(1-p)}{n}\cdot\left(1+\frac{e^{-2(p-\frac{1}{2})^2}-1}{2^k}\right)^n\right)\left(1-2^k\cdot\left(1+\frac{e^{-2(p-\frac{1}{2})^2}-1}{2^k}\right)^n\right)\\
&\quad + \frac{p^2+(1-p)^2}{2}\left(1-\frac{1}{2^k}\right)^n\\
&\geq \frac{2p(1-p)}{n} - 2^{k+2}\cdot\left(1-2^{-k}\left(1-\sqrt{4p(1-p)}\right)\right)^n - \frac{2^{k+1}p(1-p)}{n}\cdot\left(1+\frac{e^{-2(p-\frac{1}{2})^2}-1}{2^k}\right)^n\\
&\quad - \frac{2^{k+1}p(1-p)}{n}\cdot\left(1+\frac{e^{-2(p-\frac{1}{2})^2}-1}{2^k}\right)^n + \frac{p^2+(1-p)^2}{2}\left(1-\frac{1}{2^k}\right)^n\\
&\geq \frac{2p(1-p)}{n} - 2^{k+2}\cdot\left(1-2^{-k}\left(1-\sqrt{4p(1-p)}\right)\right)^n - \frac{2^{k+2}p(1-p)}{n}\cdot\left(1-\frac{1-e^{-2(p-\frac{1}{2})^2}}{2^k}\right)^n\\
&\quad + \frac{p^2+(1-p)^2}{2}\left(1-\frac{1}{2^k}\right)^n\\
&\geq \frac{2p(1-p)}{n} - \frac{2^{k+3}\cdot(1-\rho_{k,p})^n}{n} + \frac{p^2+(1-p)^2}{2}\left(1-\frac{1}{2^k}\right)^n
\end{aligned}$$

where

$$\rho_{k,p} = 2^{-k}\min\left(1-\sqrt{4p(1-p)}, 1-e^{-2(p-\frac{1}{2})^2}\right).$$

Note that, since $p > 1/2$, $0 < \rho_{k,p} < 1$.

Lemma S7. *Let S be a positive random variable. For any real-valued $\alpha \in [0,1]$, for any $n \in \mathbb{N}$,*

$$\mathbb{P}\left(S \leq \alpha n\right)\mathbb{V}[S|S\leq \alpha n] \leq \mathbb{V}[S]$$

Proof. We start by noticing that:

$$\begin{aligned}
A_n &= \mathbb{P}\left(S>\alpha n\right)\mathbb{E}\left[\left(S-\mathbb{E}\left[S \mid S>\alpha n\right]\right)^2 \mid S>\alpha n\right]\\
&\quad + \mathbb{P}\left(S\leq\alpha n\right)\mathbb{E}\left[\left(S-\mathbb{E}\left[S \mid S\leq\alpha n\right]\right)^2 \mid S\leq\alpha n\right]\\
&\leq \mathbb{P}\left(S>\alpha n\right)\mathbb{E}\left[\left(S-a\right)^2 \mid S>\alpha n\right] + \mathbb{P}\left(S\leq\alpha n\right)\mathbb{E}\left[\left(S-b\right)^2 \mid S\leq\alpha n\right]
\end{aligned}$$

for any $(a,b)\in\mathbb{R}^2$.

Then,

$$\begin{aligned}
A_n &\leq \mathbb{P}\left(S>\alpha n\right)\mathbb{E}\left[\left(S-a\right)^2 \mid S>\alpha n\right] + \mathbb{P}\left(S\leq\alpha n\right)\mathbb{E}\left[\left(S-a\right)^2 \mid S\leq\alpha n\right]\\
&= \mathbb{E}\left[\left(S-a\right)^2\right]
\end{aligned}$$

for any $a \in \mathbb{R}$.

Choosing $a = \mathbb{E}[S]$, we obtain

$$A_n \leq \mathbb{V}[S].$$

Therefore,

$$\mathbb{P}(S \leq \alpha n)\, \mathbb{V}[S \mid S \leq \alpha n] \leq \mathbb{V}[S].$$

□

S6 Extended results for a random chessboard

Proposition S8 (Risk of a single tree and a shallow tree network when $k < k^\star$)**.** *Let $N \in \{1, ..., 2^{k^\star}\}$. We consider the data distribution defined by a random chessboard with i.i.d. cells such that for each cell $C_i, i \in \{1, ..., 2^{k^\star}\}$*

$$\mathbb{P}(C_i \subset \mathcal{B}) = \frac{N}{2^{k^\star}}$$

and $\mathbb{P}(C_i \subset \mathcal{W}) = 1 - \frac{N}{2^{k^\star}}$. Notice that the (random) numbers $N_\mathcal{W}$ and $N_\mathcal{B}$ of white and black cells satisfy $0 \leq N_\mathcal{W} = 2^{k^\star} - N_\mathcal{B} \leq 2^{k^\star}$. We study the risk of the shallow tree network $\hat{r}_{k,1,n}$.

1. *Consider a single tree $\hat{r}_{k,0,n}$ of depth $k \in \mathbb{N}^\star$,*

$$R(\hat{r}_{k,0,n}) \leq 4(p - \frac{1}{2})^2 \frac{N}{2^{k^\star}} \left(1 - \frac{N}{2^{k^\star}}\right) \left(1 + \frac{1}{2^{k^\star - k}}\right) + \frac{2^{k-1}}{n+1} + \left((1-p)^2 - \frac{N}{2^{k^\star}}(1-2p)\right)(1 - 2^{-k})^n$$

and

$$R(\hat{r}_{k,0,n}) \geq 4(p - \frac{1}{2})^2 \frac{N}{2^{k^\star}} \left(1 - \frac{N}{2^{k^\star}}\right) + \frac{2^k}{n+1}(1-p)^2 + C_{k^\star,k,N,p}(1 - 2^{-k})^n$$

where $C_{k^\star,k,N,p} = (1-p)^2 - \frac{N}{2^{k^\star}}(1-2p) - \frac{(1-p)^2 2^k}{n+1} - 4(p - \frac{1}{2})^2 \frac{N}{2^{k^\star}} \left(1 - \frac{N}{2^{k^\star}}\right) \left(1 + \frac{1}{2^{k^* - k}}\right)$.

2. *Consider the shallow tree network $\hat{r}_{k,1,n}$, in the infinite sample regime,*

$$R(\hat{r}_{k,1,n}) \geq \left(p - \frac{1}{2}\right)^2 \min\left(1 - \frac{N}{2^{k^\star}}, \frac{N}{2^{k^\star}}\right)^2$$

and

$$R(\hat{r}_{k,1,n}) \leq 4\left(p - \frac{1}{2}\right)^2 \left(1 - \frac{N}{2^{k^\star}}\right) \frac{N}{2^{k^\star}} + p^2 \min\left(\frac{N}{2^{k^\star}}, 1 - \frac{N}{2^{k^\star}}\right).$$

S7 Proof of Proposition S8

S7.1 First statement: risk of a single tree

To see the definitions of $\tilde{\mathcal{B}}$ and $\tilde{\mathcal{W}}$ refer to the notations of the second statement of the proof of Proposition 3, in Appendix S5.2.

Recall that $k < k^\star$, meaning that a tree leaf may contain black and white cells. If a cell is empty, the tree prediction in this cell is set (arbitrarily) to zero. Thus,

$$\begin{aligned}
&\mathbb{E}\left[(\hat{r}_{k,0,n}(X) - r(X))^2\right] \\
&= \mathbb{E}\left[(\hat{r}_{k,0,n}(X) - r(X))^2 \mathbb{1}_{N_n(L_n(X))>0}\right] + \mathbb{E}\left[(r(X))^2 \mathbb{1}_{N_n(L_n(X))=0}\right] \qquad (92) \\
&= \mathbb{E}\left[\left(\frac{1}{N_n(L_n(X))} \sum_{X_i \in L_n(X)} Y_i - r(X)\right)^2 \mathbb{1}_{N_n(L_n(X))>0}\right] + \mathbb{E}\left[(r(X))^2 \mathbb{1}_{N_n(L_n(X))=0}\right], \qquad (93)
\end{aligned}$$

where the expectation is taken over the distribution of the chessboard, (X, Y) and the dataset $(X_i, Y_i)_{1\leq i\leq n}$. Besides,

$$\mathbb{E}\left[(r(X))^2 \mathbb{1}_{N_n(L_n(X))=0}\right] = \mathbb{E}\left[(r(X))^2 \mathbb{1}_{N_n(L_n(X))=0}\mathbb{1}_{X\in\mathcal{B}}\right] + \mathbb{E}\left[(r(X))^2 \mathbb{1}_{N_n(L_n(X))=0}\mathbb{1}_{X\in\mathcal{W}}\right] \quad (94)$$

$$= ((1-p)^2 - \frac{N_{\mathcal{B}}}{2^{k^\star}}(1-2p))(1-2^{-k})^n \quad (95)$$

We now study the first term in (93), by considering that X falls into $\mathcal{B}$ (the same computation holds when X falls into $\mathcal{W}$). We denote $|L_n(X)\cap\mathcal{B}||$ (resp. $|L_n(X)\cap\mathcal{W}|$) the number of black (resp. white) cells included in the cell containing X. Letting (X', Y') generic random variables with the same distribution as (X, Y), one has

$$\mathbb{E}\left[\left(\frac{1}{N_n(L_n(X))}\sum_{X_i\in L_n(X)} Y_i - p\right)^2 \mathbb{1}_{N_n(L_n(X))>0}\mathbb{1}_{X\in\mathcal{B}}\right] \quad (96)$$

$$= \mathbb{P}(X\in\mathcal{B})\,\mathbb{E}\left[\left(\frac{1}{N_n(L_n(X))}\sum_{X_i\in L_n(X)} (Y_i - \mathbb{E}[Y'|X'\in L_n(X), |L_n(X)\cap\mathcal{B}|])\right)^2 \mathbb{1}_{N_n(L_n(X))>0}\right] \quad (97)$$

$$+ \mathbb{P}(X\in\mathcal{B})\,\mathbb{E}\left[\left(\mathbb{E}[Y'|X'\in L_n(X), |L_n(X)\cap\mathcal{B}|] - p\right)^2 \mathbb{1}_{N_n(L_n(X))>0}\right]$$

$$= \mathbb{P}(X\in\mathcal{B})\cdot$$

$$\mathbb{E}\left[\frac{\mathbb{1}_{N_n(L_n(X))>0}}{N_n(L_n(X))^2}\mathbb{E}\left[\left(\sum_{X_i\in L_n(X)} (Y_i - \mathbb{E}[Y'|X'\in L_n(X), |L_n(X)\cap\mathcal{B}|])\right)^2 \mid N_n(L_n(X)), |L_n(X)\cap\mathcal{B}|\right]\right] \quad (98)$$

$$+ \mathbb{P}(X\in\mathcal{B})\,\mathbb{P}(N_n(L_n(X))>0)\,\mathbb{E}\left[\left(\mathbb{E}[Y'|X'\in L_n(X), |L_n(X)\cap\mathcal{B}|] - p\right)^2\right]$$

$$= \mathbb{P}(X\in\mathcal{B})\left(\mathbb{E}\left[\frac{\mathbb{1}_{N_n(L_n(X))>0}}{N_n(L_n(X))^2}\cdot V_{\mathcal{B}}\right] + \beta_{\mathcal{B}}\right) \quad (99)$$

where

$$\beta_{\mathcal{B}} = \mathbb{P}(N_n(L_n(X))>0)\,\mathbb{E}\left[\left(\mathbb{E}[Y'|X'\in L_n(X), |L_n(X)\cap\mathcal{B}|] - p\right)^2\right] \quad (100)$$

and

$$V_{\mathcal{B}} = \mathbb{E}\left[\left(\sum_{X_i\in L_n(X)} (Y_i - \mathbb{E}[Y'|X'\in L_n(X), |L_n(X)\cap\mathcal{B}|])\right)^2 \mid N_n(L_n(X)), |L_n(X)\cap\mathcal{B}|\right] \quad (101)$$

Similarly we define β_W and $V_{\mathcal{W}}$ by replacing in the expressions (100) and (101) $\mathcal{B}$ by $\mathcal{W}$ so that:

$$\mathbb{E}\left[(\hat{r}_{k,0,n}(X) - r(X))^2\right]$$
$$= \mathbb{P}(X\in\mathcal{B})\left(\mathbb{E}\left[\frac{\mathbb{1}_{N_n(L_n(X))>0}}{N_n(L_n(X))^2}\cdot V_{\mathcal{B}}\right] + \beta_{\mathcal{B}}\right) + \mathbb{P}(X\in\mathcal{W})\left(\mathbb{E}\left[\frac{\mathbb{1}_{N_n(L_n(X))>0}}{N_n(L_n(X))^2}\cdot V_{\mathcal{W}}\right] + \beta_{\mathcal{W}}\right)$$
$$+ ((1-p)^2 - \frac{N_{\mathcal{B}}}{2^{k^\star}}(1-2p))(1-2^{-k})^n. \quad (102)$$

This last expression can be read as a bias-variance decomposition. We already know the probability to be in a non-empty cell, see (95), then

$$\mathbb{P}(X\in\mathcal{B}) = \mathbb{E}[\mathbb{P}(X\in\mathcal{B}|N_{\mathcal{B}})] = \mathbb{E}\left[\frac{N_{\mathcal{B}}}{2^{k^\star}}\right] = \frac{N}{2^{k^\star}}.$$

We now make explicit the terms in Equation (102) starting with the bias term $\beta_{\mathcal{B}}$:

$$\mathbb{E}\left[Y'|X' \in L_n(X), |L_n(X) \cap \mathcal{B}|\right] = \frac{p \cdot |L_n(X) \cap \mathcal{B}| + (1-p)|L_n(X) \cap \mathcal{W}|}{2^{k^\star - k}} \tag{103}$$

$$= (1-p) + \frac{|L_n(X) \cap \mathcal{B}|}{2^{k^\star - k}}(2p-1) \tag{104}$$

$$= p + \frac{|L_n(X) \cap \mathcal{W}|}{2^{k^\star - k}}(1-2p), \tag{105}$$

where $|L_n(X) \cap \mathcal{B}|$ stands for the number of black cells in $L_n(X)$. In the same way, $|L_n(X) \cap \mathcal{W}|$ stands for the number of white cells in $L_n(X)$. Hence,

$$\mathbb{E}\left[\left(\mathbb{E}\left[Y'|X' \in L_n(X), |L_n(X) \cap \mathcal{B}|\right] - p\right)^2\right] = \frac{4(p-\frac{1}{2})^2}{2^{2(k^\star - k)}}\mathbb{E}\left[|L_n(X) \cap \mathcal{W}|^2\right] \tag{106}$$

Note that $|L_n(X) \cap \mathcal{W}||N_{\mathcal{B}} \sim \mathfrak{B}(2^{k^\star - k}, 1 - N/2^{k^\star})$. Thus, we have

$$\mathbb{E}\left[|L_n(X) \cap \mathcal{W}|^2\right] = \frac{N}{2^k}\left(1 - \frac{N}{2^{k^\star}}\right) + \frac{1}{2^{2k}}(2^{k^\star} - N)^2. \tag{107}$$

Therefore,

$$\mathbb{E}\left[\left(\mathbb{E}\left[Y'|X' \in L_n(X), |L_n(X) \cap \mathcal{B}|\right] - p\right)^2\right] = \frac{4(p-\frac{1}{2})^2}{2^{2(k^\star - k)}}\left(\frac{N}{2^k}\left(1 - \frac{N}{2^{k^\star}}\right) + \frac{1}{2^{2k}}(2^{k^\star} - N)^2\right). \tag{108}$$

Similar computations show that when $X \in \mathcal{W}$,

$$\mathbb{E}\left[\left(\mathbb{E}\left[Y'|X' \in L_n(X), |L_n(X) \cap \mathcal{W}|\right] - (1-p)\right)^2\right] = \frac{4(p-\frac{1}{2})^2}{2^{2(k^\star - k)}}\left(\frac{N}{2^k}\left(1 - \frac{N}{2^{k^\star}}\right) + \frac{N^2}{2^{2k}}\right). \tag{109}$$

We deduce from Equations (108) and (109) that

$$\begin{aligned}\beta_{\mathcal{B}} + \beta_{\mathcal{W}} = \frac{4(p-\frac{1}{2})^2}{2^{2(k^\star - k)}}\mathbb{P}\left(N_n(L_n(X)) > 0\right)\Bigg(&\mathbb{P}\left(X \in \mathcal{B}\right)\left(\frac{N}{2^k}\left(1 - \frac{N}{2^{k^\star}}\right) + \frac{1}{2^{2k}}(2^{k^\star} - N)^2\right) \\ &+ \mathbb{P}\left(X \in \mathcal{W}\right)\left(\frac{N}{2^k}\left(1 - \frac{N}{2^{k^\star}}\right) + \frac{N^2}{2^{2k}}\right)\Bigg) \\ = 4(p-\frac{1}{2})^2\frac{N}{2^{k^\star}}&\left(1 - \frac{N}{2^{k^\star}}\right)\left(1 + \frac{1}{2^{k^\star - k}}\right)\left(1 - (1-2^{-k})^n\right).\end{aligned} \tag{110}$$

Clearly,

$$\beta_{\mathcal{B}} + \beta_{\mathcal{W}} \geq 4(p-\frac{1}{2})^2\frac{N}{2^{k^\star}}\left(1 - \frac{N}{2^{k^\star}}\right)\left(1 - (1-2^{-k})^n\right). \tag{111}$$

Now we compute the variance term $V_{\mathcal{B}}$. Letting $Z = \sum_{X_i \in L_n(X)} Y_i$,

$$Z|N_n(L_n(X)), |L_n(X) \cap \mathcal{B}| \sim \mathfrak{B}(N_n(L_n(X)), p')$$

where $p' = (1-p) + \frac{|L_n(X) \cap \mathcal{B}|}{2^{k^\star - k}}(2p-1)$ (see Equations (103) to (105)). Therefore, recall that $V_{\mathcal{B}}$ is nothing but the variance of the binomial random variable Z conditional on $|L_n(X) \cap \mathcal{B}|$ defined in Equation (101), consequently

$$V_{\mathcal{B}} = N_n(L_n(X))p'(1-p'). \tag{112}$$

By independence of $N_n(L_n(X))$ and $|L_n(X) \cap \mathcal{B}|$, we can write that

$$\mathbb{E}\left[\frac{\mathbb{1}_{N_n(L_n(X))>0}}{N_n(L_n(X))^2} \cdot V_{\mathcal{B}}\right] = \mathbb{E}\left[\frac{\mathbb{1}_{N_n(L_n(X))>0}}{N_n(L_n(X))}\right]\mathbb{E}\left[\underbrace{p'(1-p)'}_{(1-p)^2 \leq p'(1-p') \leq 1/4}\right]. \tag{113}$$

From Technical Lemma S5, we deduce that

$$\frac{2^k}{n+1}\left(1-(1-2^{-k})^n\right) \leq \mathbb{E}\left[\frac{\mathbb{1}_{N_n(L_n(X))>0}}{N_n(L_n(X))}\right] \leq \frac{2^{k+1}}{n+1}.$$

Hence,

$$\frac{2^k}{n+1}\left(1-(1-2^{-k})^n\right)(1-p)^2 \leq \mathbb{E}\left[\frac{\mathbb{1}_{N_n(L_n(X))>0}}{N_n(L_n(X))}V_{\mathcal{B}}\right] \leq \frac{2^{k-1}}{n+1}. \tag{114}$$

By symmetry, $V_{\mathcal{W}}$ is also the variance of a binomial random variable with parameters $N_n(L_n(X))$, $1-p'$ conditional on $|L_n(X)\cap\mathcal{W}|$. Thus $V_{\mathcal{B}} = V_{\mathcal{W}}$. To conclude, combining Equations (102), (111) and (114) leads to

$$R(\hat{r}_{k,0,n}(X)) \leq 4(p-\frac{1}{2})^2\left(1-\frac{N}{2^{k^\star}}(1-\frac{N}{2^{k^\star}})\right)+\frac{2^{k-1}}{n+1}+((1-p)^2-\frac{N}{2^{k^\star}}(1-2p))(1-2^{-k})^n$$

and

$$\begin{aligned}R(\hat{r}_{k,0,n}(X)) \geq 4(p-\frac{1}{2})^2\left(1-\frac{N}{2^{k^\star}}(2-\frac{N}{2^{k^\star}})\right)\frac{2^k}{n+1}\left(1-(1-2^{-k})^n\right)(1-p)^2\\ +((1-p)^2-\frac{N}{2^{k^\star}}(1-2p))(1-2^{-k})^n.\end{aligned}$$

S7.2 Second statement: risk of a shallow tree network

Recall that we are in the infinite sample regime and that $k < k^\star$.

$$\begin{aligned}\mathbb{E}\left[(\hat{r}_{k,1,n}(X)-r(X)^2\right] = \mathbb{E}\left[(\hat{r}_{k,1,n}(X)-p)^2\mathbb{1}_{X\in\mathcal{B}\cap\tilde{\mathcal{B}}}\right]+\mathbb{E}\left[(\hat{r}_{k,1,n}(X)-(1-p))^2\mathbb{1}_{X\in\mathcal{W}\cap\tilde{\mathcal{W}}}\right]\\ +\mathbb{E}\left[(\hat{r}_{k,1,n}(X)-(1-p))^2\mathbb{1}_{X\in\mathcal{W}\cap\tilde{\mathcal{B}}}\right]+\mathbb{E}\left[(\hat{r}_{k,1,n}(X)-p)^2\mathbb{1}_{X\in\mathcal{B}\cap\tilde{\mathcal{W}}}\right].\end{aligned} \tag{115}$$

We begin with the computation of the first term.

$$\mathbb{E}\left[(\hat{r}_{k,1,n}(X)-p)^2\mathbb{1}_{X\in\mathcal{B}\cap\tilde{\mathcal{B}}}\right] = \mathbb{P}\left(X\in\mathcal{B}\cap\tilde{\mathcal{B}}\right)\mathbb{E}\left[(\hat{r}_{k,1,n}(X)-p)^2 \mid X\in\mathcal{B}\cap\tilde{\mathcal{B}}\right] \tag{116}$$

$$= \mathbb{P}\left(X\in\mathcal{B}\cap\tilde{\mathcal{B}}\right)\mathbb{E}\left[\left(\mathbb{E}\left[Y' \mid X'\in\tilde{\mathcal{B}}\right]-p\right)^2 \mid X\in\mathcal{B}\cap\tilde{\mathcal{B}}\right]. \tag{117}$$

Regarding the probability term,

$$\mathbb{P}\left(X\in\mathcal{B}\cap\tilde{\mathcal{B}}\right) = \mathbb{P}\left(X\in\mathcal{B}\right)\mathbb{P}\left(X\in\tilde{\mathcal{B}} \mid X\in\mathcal{B}\right) \tag{118}$$

$$\leq \frac{N}{2^{k^\star}}. \tag{119}$$

We denote by $B_1,...,B_{2^k}$ the number of black cells in the leaves $L_1,...,L_{2^k}$. Then,

$$\begin{aligned}\mathbb{E}\left[Y' \mid X'\in\tilde{\mathcal{B}}\right] &= \mathbb{E}\left[(1-p)+(2p-1)\frac{\sum_{i=1}^{2^k}B_i\mathbb{1}_{L_i\subset\tilde{\mathcal{B}}}}{|\tilde{\mathcal{B}}|}\right]\\ &= (1-p)+(2p-1)\mathbb{E}\left[\sum_{i=1}^{2^k}\frac{\mathbb{1}_{L_i\subset\tilde{\mathcal{B}}}}{|\tilde{\mathcal{B}}|}\mathbb{E}\left[B_i\middle||\tilde{\mathcal{B}}|\right]\right]\\ &= (1-p)+(2p-1)\mathbb{E}\left[\frac{B_j}{2^{k^\star-k}} \mid B_j\geq\frac{|L_j|}{2}\right]\end{aligned}$$

where L_j is a leaf included in $\tilde{\mathcal{B}}$. Moreover,

$$\begin{aligned}\mathbb{E}\left[B_j \mid B_j \geq \frac{|L_j|}{2}\right] &= \frac{N}{2^k} + \left(1 - \frac{N}{2^{k^\star}}\right)\left(2^{k^\star - k - 1} - 1\right)\frac{\mathbb{P}\left(B_j = 2^{k^\star - k - 1} - 1\right)}{\mathbb{P}\left(B_j \geq 2^{k^\star - k - 1} - 1\right)} \\ &\leq \frac{N}{2^k} + \left(1 - \frac{N}{2^{k^\star}}\right)2^{k^\star - k - 1}.\end{aligned}$$

Therefore,

$$\mathbb{E}\left[Y' \mid X' \in \tilde{\mathcal{B}}\right] \leq (1-p) + (2p-1)\frac{1}{2}\left(1 + \frac{N}{2^{k^\star}}\right)$$

and

$$\mathbb{E}\left[\left(\mathbb{E}\left[Y' \mid X' \in \tilde{\mathcal{B}}\right] - p\right)^2 \mid X \in \mathcal{B} \cap \tilde{\mathcal{B}}\right] \geq (p - \frac{1}{2})^2\left(1 - \frac{N}{2^{k^\star}}\right)^2. \tag{120}$$

To compute the upper bound, note that

$$\mathbb{E}\left[B_j \mid B_j \geq \frac{|L_j|}{2}\right] \geq \mathbb{E}\left[B_j\right] = \frac{N}{2^k}.$$

Thus,

$$\mathbb{E}\left[\left(\mathbb{E}\left[Y' \mid X' \in \tilde{\mathcal{B}}\right] - p\right)^2 \mid X \in \mathcal{B} \cap \tilde{\mathcal{B}}\right] \leq 4\left(p - \frac{1}{2}\right)^2\left(1 - \frac{N}{2^{k^\star}}\right)^2. \tag{121}$$

We adapt the previous computations to the term $\mathbb{E}\left[(\hat{r}_{k,1,n}(X) - (1-p))^2 \mathbb{1}_{X \in \mathcal{W} \cap \tilde{\mathcal{W}}}\right]$ from Equation (115). We have

$$\mathbb{E}\left[(\hat{r}_{k,1,n}(X) - (1-p))^2 \mid X \in \mathcal{W} \cap \tilde{\mathcal{W}}\right] \geq (p - \frac{1}{2})^2 \frac{N^2}{2^{2k^\star}} \tag{122}$$

and

$$\mathbb{E}\left[(\hat{r}_{k,1,n}(X) - (1-p))^2 \mid X \in \mathcal{W} \cap \tilde{\mathcal{W}}\right] \leq 4\left(p - \frac{1}{2}\right)^2 \frac{N^2}{2^{2k^\star}} \tag{123}$$

Moreover, note that

$$\mathbb{E}\left[(\hat{r}_{k,1,n}(X) - p)^2 \mathbb{1}_{X \in \mathcal{B} \cap \tilde{\mathcal{W}}}\right] \leq p^2 \mathbb{P}\left(X \in \mathcal{B} \cap \tilde{\mathcal{W}}\right) \tag{124}$$

and

$$\mathbb{E}\left[(\hat{r}_{k,1,n}(X) - p)^2 \mathbb{1}_{X \in \mathcal{B} \cap \tilde{\mathcal{W}}}\right] \geq \left(p - \frac{1}{2}\right)^2 \mathbb{P}\left(X \in \mathcal{B} \cap \tilde{\mathcal{W}}\right). \tag{125}$$

Similarly,

$$\mathbb{E}\left[(\hat{r}_{k,1,n}(X) - p)^2 \mathbb{1}_{X \in \mathcal{W} \cap \tilde{\mathcal{B}}}\right] \leq p^2 \mathbb{P}\left(X \in \mathcal{W} \cap \tilde{\mathcal{B}}\right) \tag{126}$$

and

$$\mathbb{E}\left[(\hat{r}_{k,1,n}(X) - p)^2 \mathbb{1}_{X \in \mathcal{W} \cap \tilde{\mathcal{B}}}\right] \geq \left(p - \frac{1}{2}\right)^2 \mathbb{P}\left(X \in \mathcal{W} \cap \tilde{\mathcal{B}}\right). \tag{127}$$

Gathering Equation (115) and Equations (120) to (127) yields

$$\begin{aligned}\mathbb{E}\left[(\hat{r}_{k,1,n}(X) - r(X)^2\right] &\geq (p - \frac{1}{2})^2\left(1 - \frac{N}{2^{k^\star}}\right)^2 \mathbb{P}\left(X \in \mathcal{B} \cap \tilde{\mathcal{B}}\right) + (p - \frac{1}{2})^2 \frac{N^2}{2^{2k^\star}} \mathbb{P}\left(X \in \mathcal{W} \cap \tilde{\mathcal{W}}\right) \\ &\quad + \left(p - \frac{1}{2}\right)^2 \mathbb{P}\left(X \in \mathcal{W} \cap \tilde{\mathcal{B}}\right) + \left(p - \frac{1}{2}\right)^2 \mathbb{P}\left(X \in \mathcal{B} \cap \tilde{\mathcal{W}}\right) \\ &\geq \left(p - \frac{1}{2}\right)^2 \min\left(1 - \frac{N}{2^{k^\star}}, \frac{N}{2^{k^\star}}\right)^2\end{aligned}$$

as well as

$$
\begin{aligned}
&\mathbb{E}\left[(\hat{r}_{k,1,n}(X) - r(X)^2\right] \\
&\leq 4\left(p-\frac{1}{2}\right)^2\left(1-\frac{N}{2^{k^\star}}\right)^2\mathbb{P}\left(X\in\mathcal{B}\cap\tilde{\mathcal{B}}\right) + 4\left(p-\frac{1}{2}\right)^2\frac{N^2}{2^{2k^\star}}\mathbb{P}\left(X\in\mathcal{W}\cap\tilde{\mathcal{W}}\right) \\
&\quad + p^2\mathbb{P}\left(X\in\mathcal{W}\cap\tilde{\mathcal{B}}\right) + p^2\mathbb{P}\left(X\in\mathcal{B}\cap\tilde{\mathcal{W}}\right) \\
&\leq 4\left(p-\frac{1}{2}\right)^2\left(1-\frac{N}{2^{k^\star}}\right)^2\frac{N}{2^{k^\star}}\mathbb{P}\left(X\in\tilde{\mathcal{B}}\mid X\in\mathcal{B}\right) + 4\left(p-\frac{1}{2}\right)^2\frac{N^2}{2^{2k^\star}}\left(1-\frac{N}{2^{k^\star}}\right)\mathbb{P}\left(X\in\tilde{\mathcal{W}}\mid X\in\mathcal{W}\right) \\
&\quad + p^2\left(1-\frac{N}{2^{k^\star}}\right)\mathbb{P}\left(X\in\tilde{\mathcal{B}}\mid X\in\mathcal{W}\right) + p^2\frac{N}{2^{k^\star}}\mathbb{P}\left(X\in\tilde{\mathcal{W}}\mid X\in\mathcal{B}\right) \\
&\leq 4\left(p-\frac{1}{2}\right)^2\left(1-\frac{N}{2^{k^\star}}\right)^2\frac{N}{2^{k^\star}} + 4\left(p-\frac{1}{2}\right)^2\frac{N^2}{2^{2k^\star}}\left(1-\frac{N}{2^{k^\star}}\right) \\
&\quad + p^2\left(1-\frac{N}{2^{k^\star}}\right)\mathbb{P}\left(X\in\tilde{\mathcal{B}}\right) + p^2\frac{N}{2^{k^\star}}\mathbb{P}\left(X\in\tilde{\mathcal{W}}\right) \\
&\leq 4\left(p-\frac{1}{2}\right)^2\left(1-\frac{N}{2^{k^\star}}\right)\frac{N}{2^{k^\star}} + p^2\max\left(\frac{N}{2^{k^\star}}1-\frac{N}{2^{k^\star}}\right).
\end{aligned}
$$