# 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

October 27, 2020

**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) [25]. In this paper, we investigate the mechanisms at work in DF and outline that DF architecture can generally be simplified into more simple and computationally efficient shallow forests networks. Despite some instability, the latter may outperform standard predictive tree-based methods. In order to precisely quantify the improvement achieved by these light network configurations over standard tree learners, we theoretically study the performance of a shallow tree network made of two layers, each one composed of a single centered tree. 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 [24] and difficult to interpret [14].

Recently, several attempts have been made to consider networks with non-differentiable modules. Among them the Deep Forest (DF) algorithm [25], 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 [13], medical imaging [19], drug interactions [18, 22] or even fraud detection [23].

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 performance 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.

**Related Works.** Different manners of stacking trees exist, as the Forwarding Thinking Deep Random Forest (FTDRF), proposed by [15], 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 [8] 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. [16] 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 [10] (implementing layers of many forests with few trees), or even on the number of features passed between two consecutive layers [18] 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 [11] 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 [20, 21], replacing the forest by more complex estimators (cascade of ExtraTrees) [2], or by combining several of the previous modifications notably incorporating data preprocessing [9]. 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 (on simulated and real-world datasets). 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 most, 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 at `https://github.com/Ludovic-arnould/Deep-Forest`.

# 2 Deep Forests

## 2.1 Description

Deep Forest [25] 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) [25] 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.

## 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 [25], 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` [17] 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,

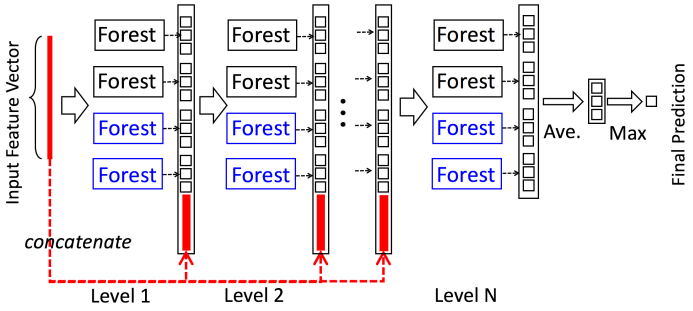



Figure 1: Deep Forest architecture (the scheme is taken from [25])

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 in the Related works paragraph, 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.** *Deep Forest [25] was first designed to handle images. To do so, a pre-processing network called Multi Grained Scanning (MGS) based on convolution methods 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 | 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 used in numerical experiments. The number of classes is shown in parenthesis for the four classification datasets.

In what follows, we propose a light DF configuration. We show that, in most cases (particularly in classification), our light configuration performance is comparable to the performance of the default DF architecture of [25], 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. Finally, 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 performance of the following DF architectures on the datasets summarized in Table 1:

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

(i) the default setting of DF introduced by [25] and described in the above section,

(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 in the forests;

(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.

**Results.** The 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 S2 (in the appendix).

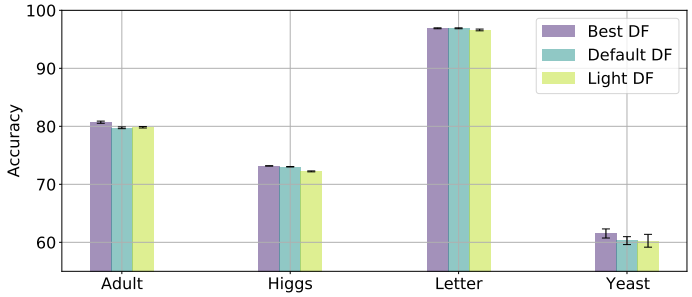



Figure 2: Comparison between different DF architectures in terms of accuracy for classification datasets (10 runs per bar plot).

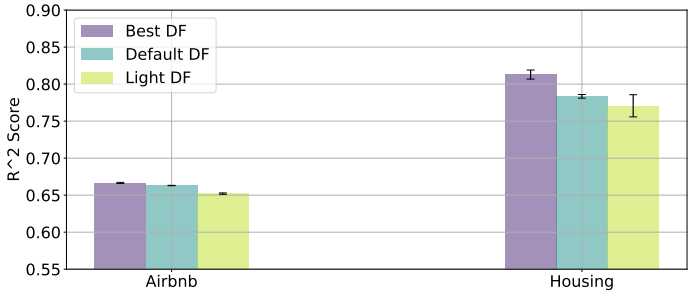



Figure 3: Comparison between different DF architectures in terms of $R^2$ score for regression datasets (10 runs per bar plot).

As highlighted in Figure 2, the 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). 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 output 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 contribution of each layer to the global DF performance.

## 3.2 Tracking the best sub-model

**Setting.** On all the previous datasets, we train a DF architecture by specifying the number $p$ of layers. Unspecified hyper-parameters are set to default value (see Section 2). For each $p$, we consider

the truncated sub-models composed of layer 1, layer 1-2, ..., layer 1-$p$, where layer 1-$p$ is the original DF with $p$ layers. For each value of $p$, we consider the previous nested sub-models with $1, 2, \ldots, p$ layers, and compute the predictive accuracy of the best sub-model.

**Results.** We only display results for the Adult dataset in Figure 4 (all the other datasets show similar results, see Appendix S1.3). We observe that adding layers to the Deep Forest does not significantly change the accuracy score. Even if the variance changes by adding layer, we are not able to detect any pattern, which suggests that the variance of the procedure performance is unstable with respect to the number of layers.

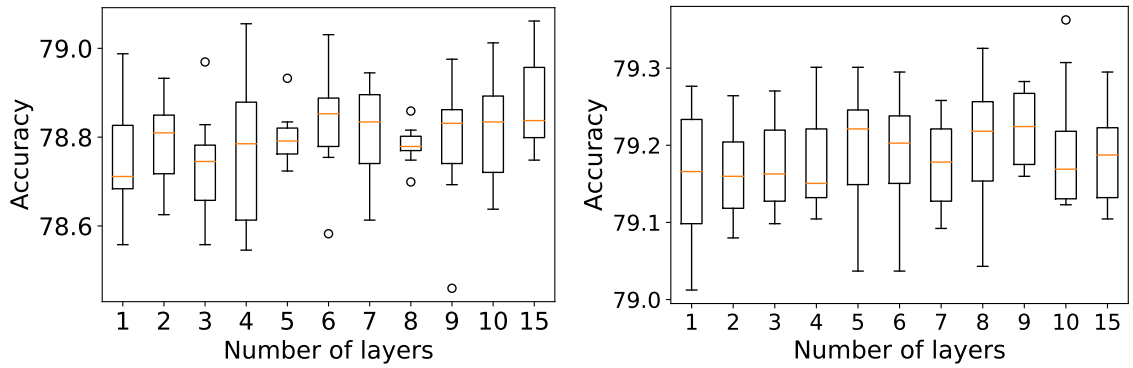



Figure 4: Adult dataset. Boxplots over 10 runs of the accuracy of a DF sub-model with 1 forest by layer (left) or 4 forests by layer (right), depending on the number of layers of the global DF model.

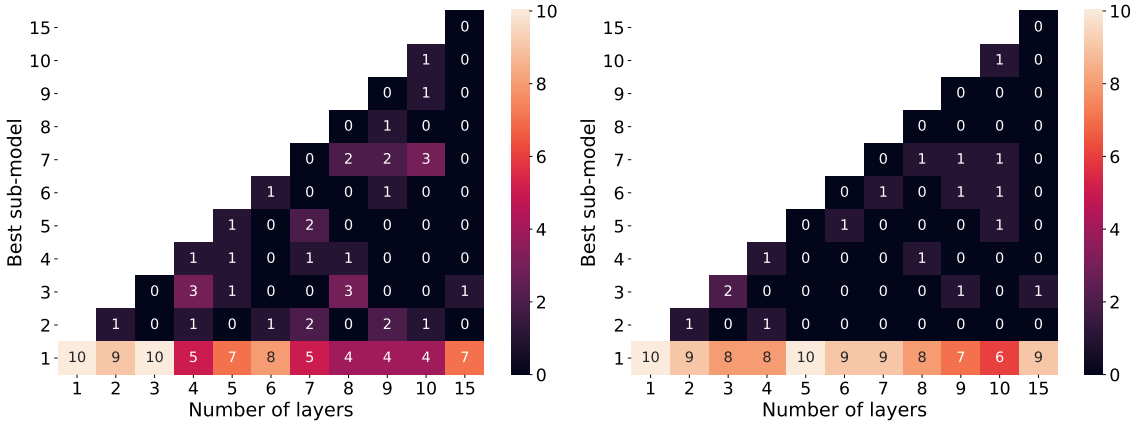



Figure 5: Adult dataset. Heatmap counting the optimal layer index over 10 tries of a default DF with 1 (Breiman) forest per layer (left) or 4 forests (2 Breiman, 2 CRF) per layer (right), with respect to the maximal number of layers. The number corresponding to $(m, n)$ on the x- and y-axes indicates how many times out of 10 the layer $m$ is optimal when running a cascade network with a number $n$ of layers.

Globally, we observe that the sub-models with one or two layers often lead to the best performance (see Figure 5 for the Adult dataset and Appendix S1.3 for the other ones). When the dataset is small (Letter or Yeast), the sub-model with only one layer (i.e. a standard RF) is almost always optimal since a single RF with no maximum depth constraint already overfits on most of these datasets. Therefore the second layer, building upon the predictions of the first layer, entails overfitting as well, therefore leading to no improvement of the overall model. Besides, one can explain the predominance of small sub-models by the weak representability power created by each layer: on the one hand, each new feature vector size corresponds to the number of classes times the number of forests which can be small with respect to the number of input features; on the other hand, the different forests within one layer are likely to produce similar probability outputs, especially if the number of trees within

each forest is large. The story is a little bit different for the Housing dataset, for which the best submodel is between 2 and 6. As noticed before, this may be the result of the frustratingly simple representation of the new features created at each layer. Eventually, these numerical experiments corroborate the relevance of shallow DF as the light configuration proposed in the previous section.

We note that adding forests in each layer decreases the number of layers needed to achieve a pre-specified performance. This is surprising and is opposed to the common belief that in deep neural networks, adding layers is usually better than adding neurons in each layer.

### 3.3 A precise understanding of tree depth in DF

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, if we consider 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 $p_0$ and $p_1$ the predicted probabilities by the first-layer tree for the classes 0 and 1 respectively.

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.

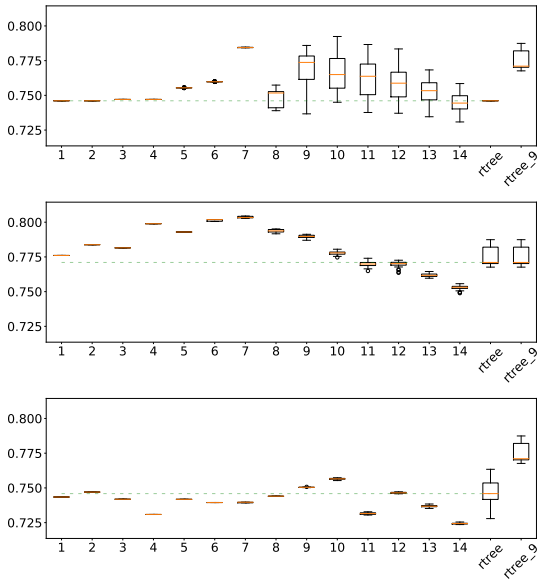



Figure 6: Adult dataset. Accuracy of a two-layer tree architecture w.r.t. the second-layer tree depth, when the first-layer (encoding) tree is of depth 2 (top), 9 (middle), and 15 (bottom). `rtree` is a single tree of respective depth 2 (top), 9 (middle), and 15 (bottom), applied on raw data. For this 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.

**Results.** Results are displayed in Figure 6 for the Adult dataset only (see Appendix S1.2 for the results on the other datasets). Specifically noticeable in Figure 6 (top), the tree network architecture can introduce performance instability when the second-layer tree grows (e.g. when the second-layer tree is successively of depth 7, 8 and 9).

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.

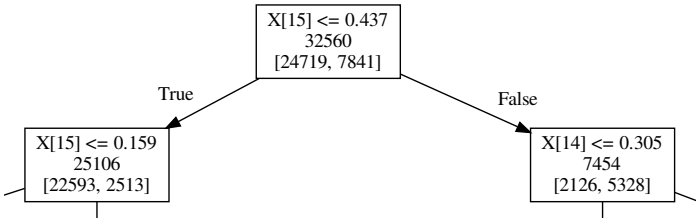



Figure 7: Adult dataset. Focus on the first levels of the second-layer tree structure when the first layer tree is of depth 9 (optimal depth). Raw features range from X[0] to X[13], X[14] and X[15] are the features built by the first-layer tree.

On all datasets, the second-layer tree is observed to always make its first cut over the new features (see Figure 7 and the ones in the Appendix S1.2 to visualize the constructed tree network structure). 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 6 (middle and bottom)). Furthermore, when considering multi-label classification with $n_{\text{classes}}$, the second-layer tree must cut over at least $n_{\text{classes}}$ features to recover the partition of the first tree (see Figure S15). 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.

In Figure 6 (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 in a binary classification setting.

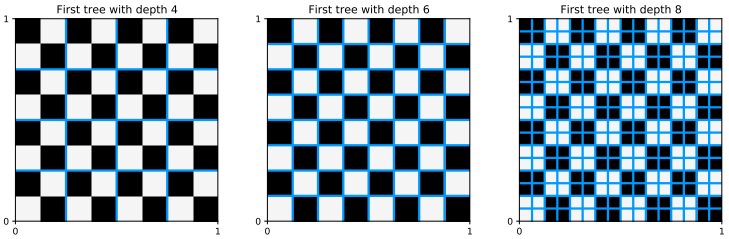
## 4.1 Problem setting

**Chessboard data generation.** Let $k^\star$ be an even integer and $p \in (1/2, 1]$. The data set $\mathcal{D}_n$ is assumed to be composed of i.i.d. pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$, with the same distribution as the generic pair $(X, Y)$. The variable $X$ is assumed to be uniformly distributed over $[0, 1]^2$ and, for all $i, j \in \{1, \ldots, 2^{k^\star/2}\}$, for all $x \in \left[\frac{i-1}{2^{k^\star/2}}, \frac{i}{2^{k^\star/2}}\right) \times \left[\frac{j-1}{2^{k^\star/2}}, \frac{j}{2^{k^\star/2}}\right)$,

$$\mathbb{P}[Y = 1 | X = x] = \begin{cases} p & \text{if } i + j \text{ is even} \\ 1 - p & \text{if } i + j \text{ is odd.} \end{cases}$$

This distribution corresponds to a chessboard structure: for each cell, which is of size $2^{-k^\star/2} \times 2^{-k^\star/2}$, either the true proportion of 1 is $p > 1/2$ or the true proportion of 0 is $p > 1/2$, depending on the parity of $i + j$ (which pinpoints the cell location). Note that the distribution is parameterized by $k^\star$ and $p$, and that $2^{k^\star}$ corresponds to the total number of cells. Such a distribution is depicted in Figure 8. This type of dataset has already been studied within RF frameworks in [5] and despite its simplicity, highlights some interesting properties of tree-based methods.

**Notations.** Given a decision tree, we will denote by $C_n(X)$ the cell of the tree containing $X$ and $N_n(C_n(X))$ the number of data points falling into $C_n(X)$. The prediction of such a tree at point $X$




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

Figure 8: Chessboard data distribution in black and white as described above for $k^\star = 6$. 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.

is given by

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

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

**A shallow centered tree network.** We want to theoretically analyze the benefits of using two trees in cascade and determine, in particular, the influence of the first (encoding) tree on the performance of the whole shallow tree network. To show the variance reduction property of the second tree already emphasized in the previous section, we need to go beyond the classical $0-1$ loss and consider instead this problem as a probability estimation one (regression setting). To this aim, we let $r(x) = \mathbb{E}[Y|X = x]$ be the regression function and we consider, for any function $f$, its quadratic risk 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 the first and second variables, at the center of the cells. The tree construction is stopped when all cells have been cut exactly $k$ times. For each point $X$, we extract the empirical mean $\bar{Y}_{C_n(X)}$ of the outputs $Y_i$ falling into the leaf $C_n(X)$ and we pass the new feature $\bar{Y}_{C_n(X)}$ 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 $\bar{Y}_{C_n(X)}$. In this latter case, two cells corresponding to $\{\bar{Y}_{C_n(X)} < 1/2\}$ and $\{\bar{Y}_{C_n(X)} \geq 1/2\}$ 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 $Y_i$ falling into a leaf containing $X$. When considering two trees in cascade, the predictor $\hat{r}_{k,k',n}(X)$ may output the mean of the $Y_i$ with the $X_i$ falling into a union of the first-tree leaves containing $X$.

## 4.2 Theoretical results

We first study the risk of the shallow tree network in the infinite sample regime. The results are presented in Lemma 1.

**Lemma 1.** *Assume that the data follows the chessboard distribution described above. In the infinite sample regime, the following holds for the shallow tree network $\hat{r}_{k,k',n}$ (Definition 1):*

*(i) For any $k < k^\star$ (shallow encoding tree), the risk of the shallow tree network is minimal for a second-layer tree of depth $k' \geq k^\star$ whose $k^\star$ first cuts are performed along raw features only.*

*(ii) For any $k \geq k^\star$ (deep encoding tree), the risk of the shallow tree network is minimal for a second-layer tree of depth $k' \geq 1$ whose first (and only) cut is performed along the new feature $\bar{Y}_{C_n(X)}$.*

The proof of Lemma 1 is given in Appendix S3. In the infinite sample regime, Lemma 1 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 1 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. We believe that in this latter case, the shallow network may benefit from the variance reduction of the second-layer tree, which gathers similar cells and averages their prediction to build the output. Indeed, this has been empirically observed when dealing with two layers of CART trees.

**Main results.** With this in mind, we move towards the finite sample regime to study the variance reduction phenomenon, and motivated by Lemma 1, we consider a second-layer tree of depth one, whose first cut is performed along the new feature $\bar{Y}_{C_n(X)}$ at $1/2$.

To study the interest of using a shallow tree network instead of a single tree, we first establish upper and lower bounds for a single centered tree of depth $k < k^\star$ and $k \geq k^\star$ respectively.

**Proposition 2** (Risk of a single tree)**.** *Assume that the data is drawn according to the chessboard distribution with parameters $k^\star$ and $p > 1/2$. Consider the predictor $\hat{r}_{k,0,n}$ corresponding to a single centered tree of depth $k \in \mathbb{N}^\star$. Then,*

*1. if $k < k^\star$,*

*(i) an upper-bound on the excess risk reads as*

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

*(ii) a lower-bound on the excess risk reads as*

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

*2. if $k \geq k^\star$,*

*(i) an upper-bound on the excess risk reads as*

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

*(ii) a lower-bound on the excess risk reads as*

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

The proof of Proposition 2 is given in Appendix S4. First, note that our bounds are tight in both cases ($k < k^\star$ and $k \geq k^\star$) since the rate of the upper bounds match that of the lower ones. The first statement in Proposition 2 quantifies the bias of a shallow 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 second statement in Proposition 2 proves that the risk of a tree deep enough ($k \geq k^\star$) tends to zero with $n$. In this case, the bias is null and the risk is governed by the variance term which is $O(2^k/n)$-term (note that $n/2^k$ is the average number of points in each cell). 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.

However, we are not interested in the performance of the single tree but in the improvements that the shallow tree network can bring to an individual tree. Note that stacking two layers of trees together still leads to a partition-type estimator with axis-aligned splits. However, it allows to build more complex partitions since it may gather cells of the first tree that are disconnected. This may lead to an improvement of the resulting estimator, by reducing the variance in the corresponding cell collections. Proposition 3 quantifies this phenomenon by establishing upper and lower bounds on the risk of the shallow tree network for $k < k^\star$ and $k \geq k^\star$.

**Proposition 3** (Risk of a shallow tree network)**.** *Assume that the data is drawn according to the chessboard distribution with parameters $k^\star$ and $p > 1/2$. Consider the predictor $\hat{r}_{k,1,n}$ corresponding to two trees in cascade (see Definition 1). Then,*

1. *if $k < k^\star$,*

   (i) *an upper-bound on the excess risk reads as*

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

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

   (ii) *a lower-bound on the excess risk reads as*

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

2. *if $k \geq k^\star$,*

   (i) *an upper-bound on the excess risk reads as*

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

   *where $\varepsilon_{n,k,p} = n\left(1 - \frac{1-e^{-2(p-\frac{1}{2})^2}}{2^k}\right)^n$.*

   (ii) *a lower-bound on the excess risk reads as*

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

   *where $0 < \rho_{k,p} < 1$ depends only on $p$ and $k$ and given that $n \geq \frac{(k+1)\log(2)}{\log(2^k)-\log(e^{-2(p-1/2)^2}-1+2^k)}$.*

The proof of Proposition 3 is given in Appendix S5. Note that, in both cases, the rate of the upper bounds match that of the lower ones, highlighting the tightness of these bounds.

As for the single tree studied in Proposition 3, the shallow tree network suffers from a bias term $(p-1/2)^2$ as soon as the first-layer tree is not deep enough. In such a shallow tree network, the flaws of the first-layer tree transfer to the whole network. However, there may exist 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, as for the chessboard distribution), the shallow tree network benefits from a variance reduction phenomenon by a factor $2^k$ (as highlighted by Proposition 3 and Proposition 2).

In Figure 9, we numerically evaluate the risk $R(\hat{r}_{k,1,n})$, and its average value exactly lies between the theoretical upper and lower bounds, that end up being merged.

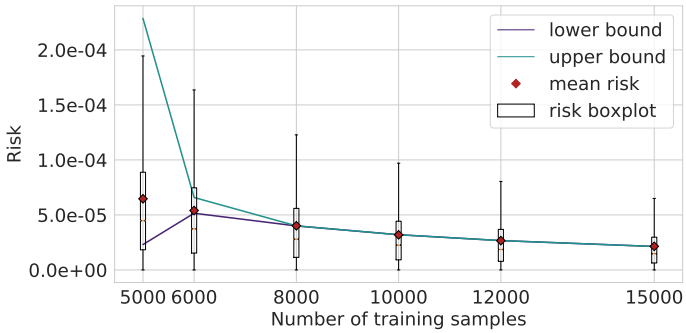



Figure 9: Illustration of the theoretical bounds of Proposition 3, when the first-layer tree is of depth $k = 6$ (when $k^\star = 4$) and $p = 0.8$. We draw a sample of size $n$ (x-axis), and a shallow tree network $r_{k,1,n}$ is fitted for which the theoretical risk is evaluated. Each boxplot is built out of 20 000 repetitions. The outliers are not shown for the sake of presentation.

# 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) closely related to DF procedure. 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, 12], 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 tree, and more generally the first layers in DF architecture, can be seen as a data-driven encoder. 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] 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.

[9] 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.

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

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

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

[13] 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.

[14] 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.

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

[16] 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.

[17] 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.

[18] 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.

[19] 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.

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

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

[22] 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.

[23] 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.

[24] 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.

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

# S1 Additional figures

## S1.1 Additional figures to Section 3.2

| 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 S2: Details of the best configurations obtained in Figures 2 and 3.

## S1.2 Additional figures to Section 3.3

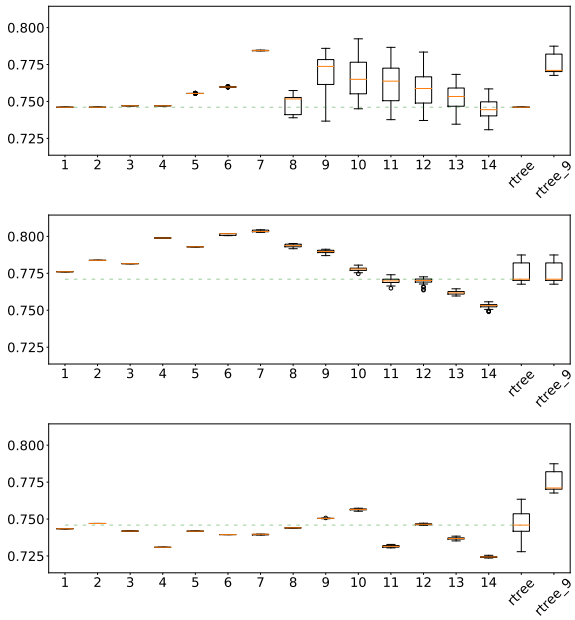



Figure S10: Adult dataset. Accuracy of a two-layer tree architecture w.r.t. the second-layer tree depth, when the first-layer (encoding) tree is of depth 2 (top), 9 (middle), and 15 (bottom). `rtree` is a single tree of respective depth 2 (top), 9 (middle), and 15 (bottom), applied on raw data. For this 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.

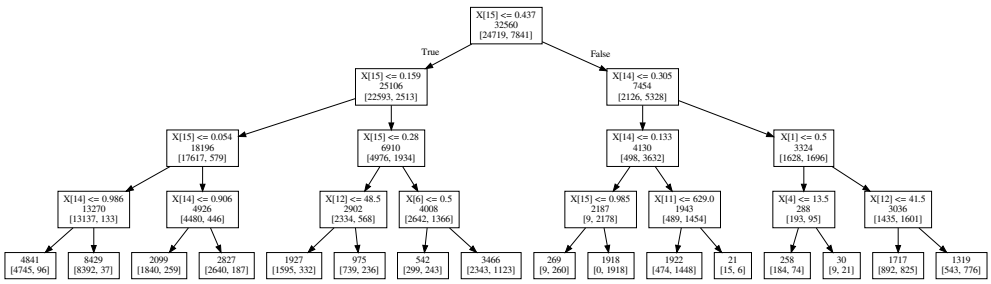



Figure S11: Adult dataset. Second-layer tree structure of depth 4 when the first-layer tree is of depth 9 (optimal depth). Raw features range from X[0] to X[13], X[14] and X[15] are the features built by the first-layer tree.

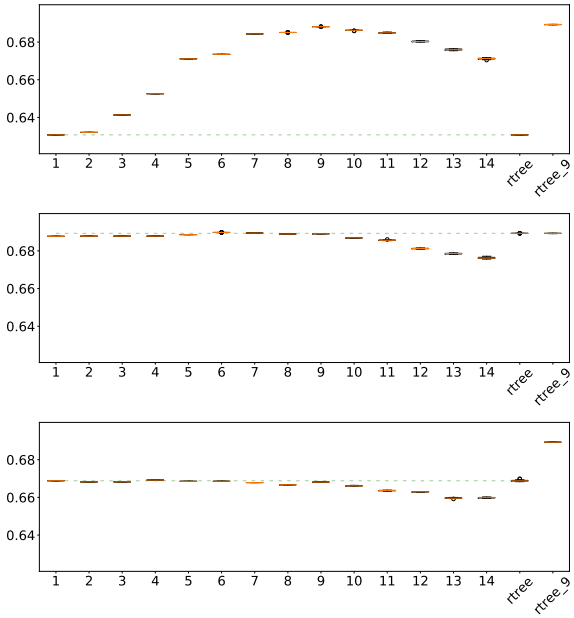



Figure S12: Higgs dataset. Accuracy of a two-layer tree architecture w.r.t. the second-layer tree depth, when the first-layer (encoding) tree is of depth 2 (top), 9 (middle), and 15 (bottom). `rtree` is a single tree of respective depth 2 (top), 9 (middle), and 15 (bottom), applied on raw data. For this 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.

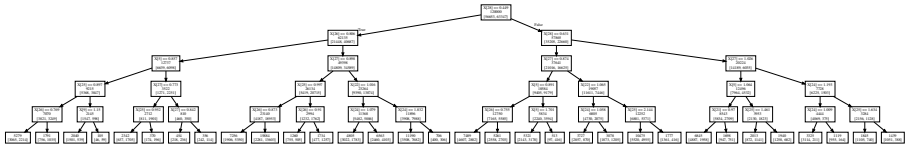


Figure S13: Higgs dataset. Second-layer tree structure of depth 5 when the first-layer tree is of depth 2 (low depth). Raw features range from X[0] to X[13], X[14] and X[15] are the features built by the first-layer tree.

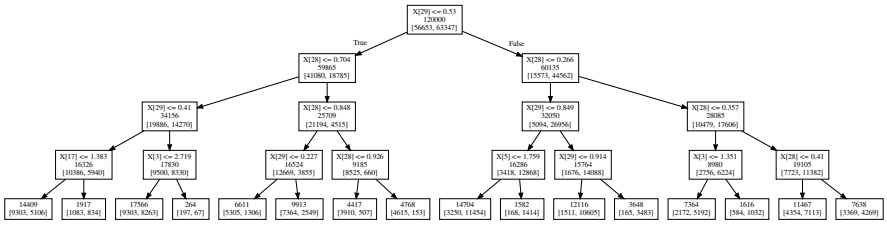



Figure S14: Higgs dataset. Second-layer tree structure of depth 4 when the first-layer tree is of depth 9 (optimal depth). Raw features range from X[0] to X[27], X[28] and X[29] are the features built by the first-layer tree.

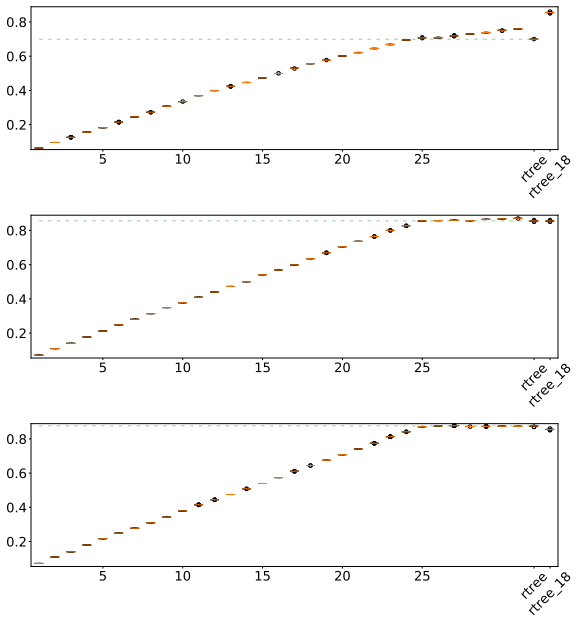



Figure S15: Letter dataset. Accuracy of a two-layer tree architecture w.r.t. the second-layer tree depth, when the first-layer (encoding) tree is of depth 10 (top), 18 (middle), and 26 (bottom). `rtree` is a single tree of respective depth 10 (top), 18 (middle), and 26 (bottom), applied on raw data. 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.

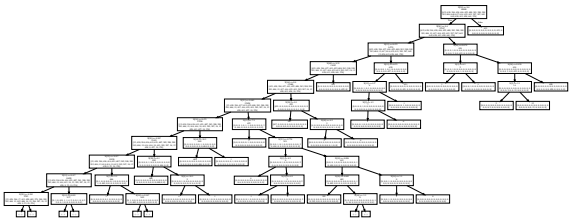


Figure S16: Letter dataset. Second-layer tree structure of depth 30 when the first-layer tree is of depth 18 (optimal depth). We only show the first part of the tree up to depth 10. Raw features range from X[0] to X[15]. The features built by the first-layer tree range from X[16] to X[41].

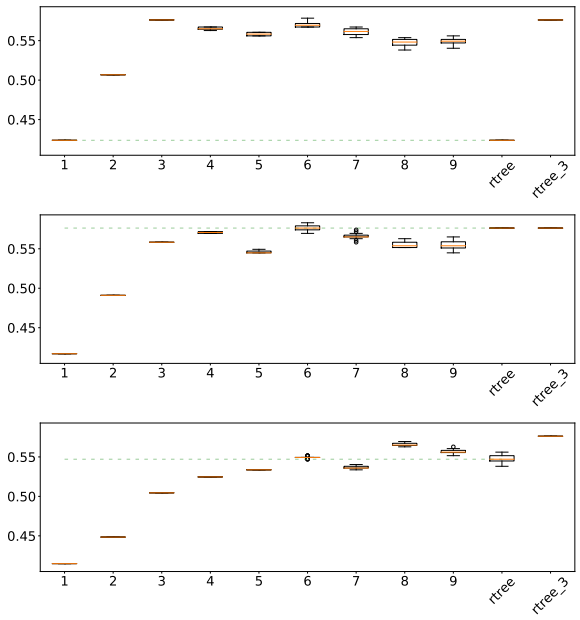



Figure S17: Yeast dataset. Accuracy of a two-layer tree architecture w.r.t. the second-layer tree depth, when the first-layer (encoding) tree is of depth 1 (top), 3 (middle), and 8 (bottom). `rtree` is a single tree of respective depth 1 (top), 3 (middle), and 8 (bottom), applied on raw data. For this 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.

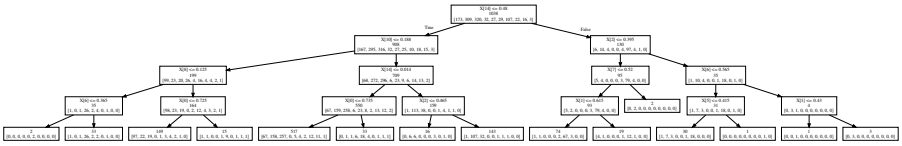


Figure S18: Yeast dataset. Second-layer tree structure of depth 4 when the first-layer tree is of depth 3 (optimal depth). Raw features range from X[0] to X[7]. The features built by the first-layer tree range from X[8] to X[17].

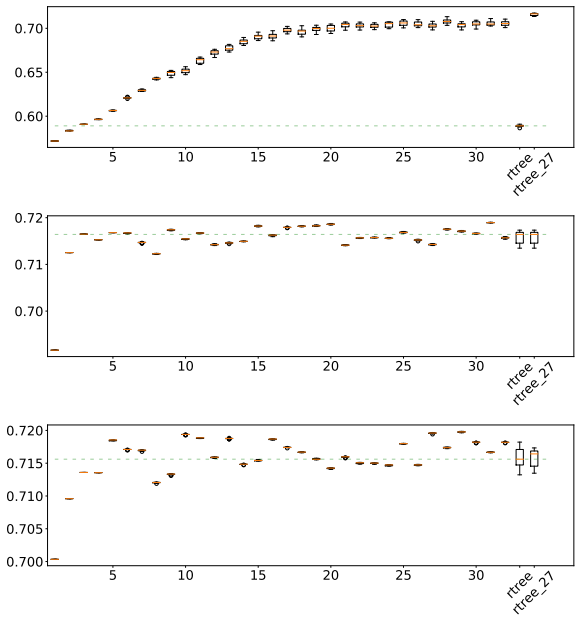



Figure S19: Airbnb dataset. $R^2$ score of a two-layer tree architecture w.r.t. the second-layer tree depth, when the first-layer (encoding) tree is of depth 10 (top), 27 (middle), and 32 (bottom). `rtree` is a single tree of respective depth 10 (top), 27 (middle), and 32 (bottom), applied on raw data. 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.

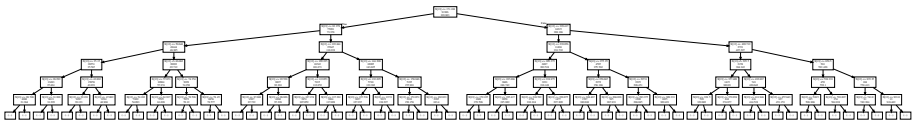


Figure S20: Airbnb dataset. Second-layer tree structure of depth 28 when the first-layer tree is of depth 26 (optimal depth). We only show the first part of the tree up to depth 5. Raw features range from X[0] to X[12], X[13] is the feature built by the first-layer tree.

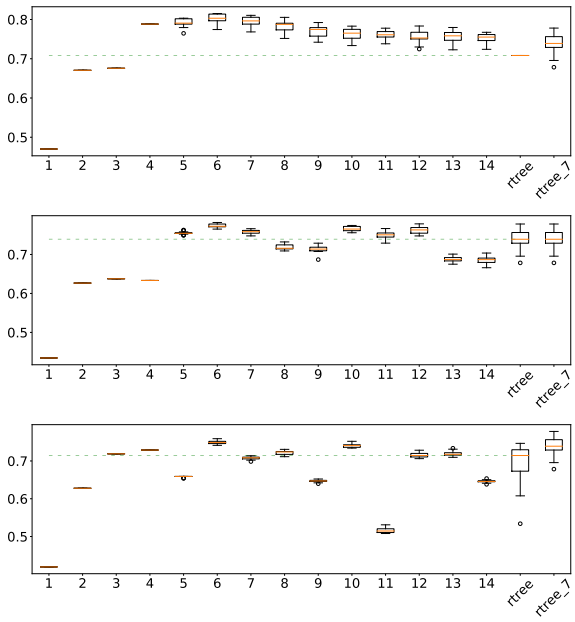



Figure S21: Housing dataset. $R^2$ score of a two-layer tree architecture w.r.t. the second-layer tree depth, when the first-layer (encoding) tree is of depth 3 (top), 7 (middle), and 12 (bottom). `rtree` is a single tree of respective depth 3 (top), 7 (middle), and 12 (bottom), applied on raw data. 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 S22: 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.

## S1.3 Additional figures to Section 3.2

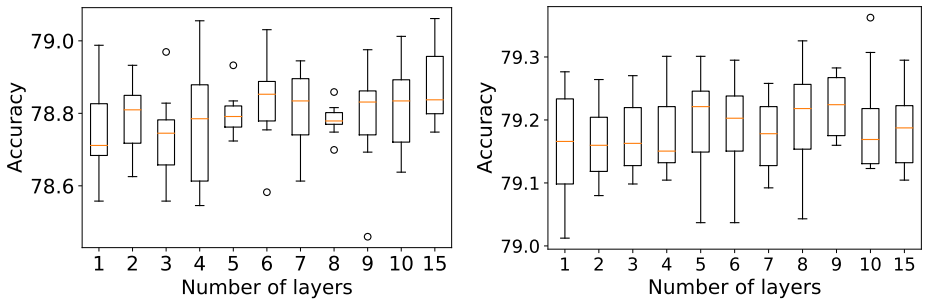



Figure S23: Adult dataset. Boxplots over 10 tries of the accuracy of a DF with 1 forest by layer (left) or 4 forests by layer (right), with respect to the DF maximal number of layers.

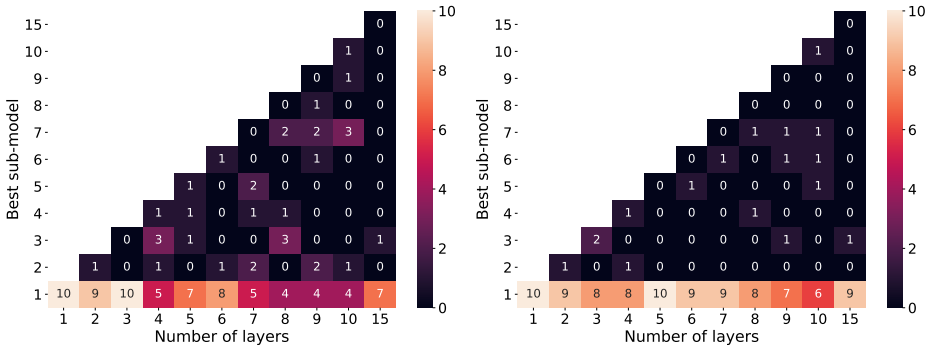



Figure S24: Adult dataset. Heatmap counting the index of the sub-optimal model over 10 tries of a default DF with 1 (Breiman) forest per layer (left) or 4 forests (2 Breiman, 2 CRF) per layer (right), with respect to the maximal number of layers.

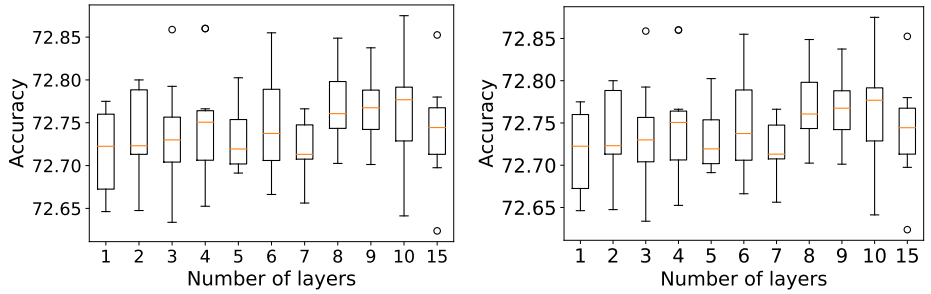



Figure S25: Higgs dataset. Boxplots over 10 tries of the accuracy of a DF with 1 forest by layer (left) or 4 forests by layer (right), with respect to the DF maximal number of layers.

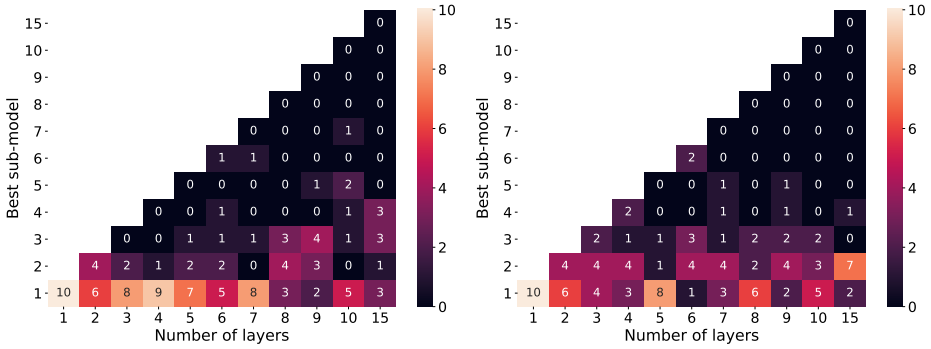



Figure S26: Higgs dataset. Heatmap counting the index of the sub-optimal model over 10 tries of a default DF with 1 (Breiman) forest per layer (left) or 4 forests (2 Breiman, 2 CRF) per layer (right), with respect to the maximal number of layers.

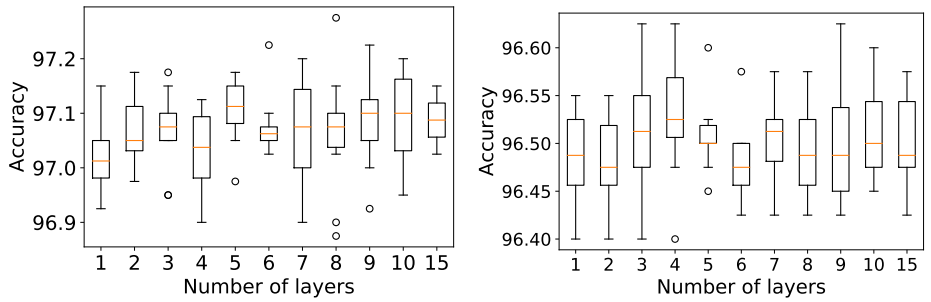



Figure S27: Letter dataset. Boxplots over 10 tries of the accuracy of a DF with 1 forest by layer (left) or 4 forests by layer (right), with respect to the DF maximal number of layers.

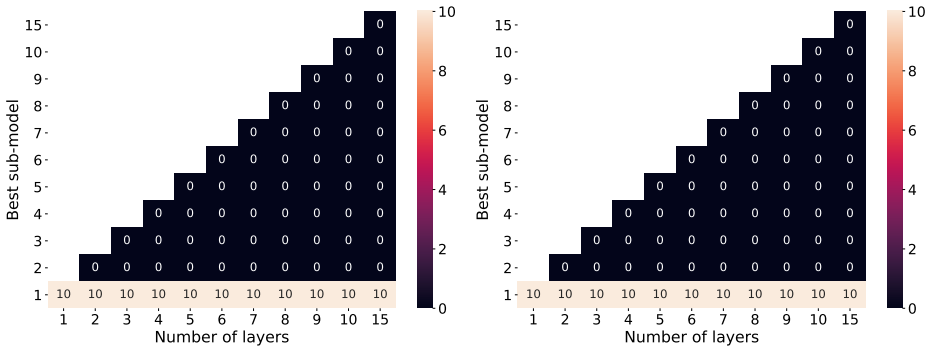



Figure S28: Letter dataset. Heatmap counting the index of the sub-optimal model over 10 tries of a default DF with 1 (Breiman) forest per layer (left) or 4 forests (2 Breiman, 2 CRF) per layer (right), with respect to the maximal number of layers.

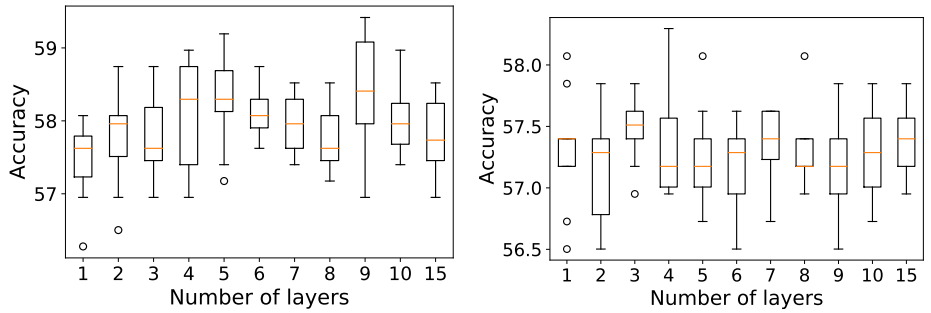



Figure S29: Yeast dataset. Boxplots over 10 tries of the accuracy of a DF with 1 forest by layer (left) or 4 forests by layer (right), with respect to the DF maximal number of layers.

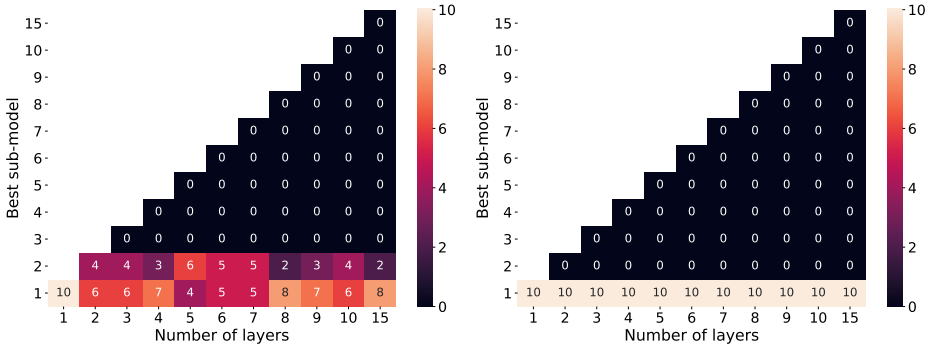



Figure S30: Yeast dataset. Heatmap counting the index of the sub-optimal model over 10 tries of a default DF with 1 (Breiman) forest per layer (left) or 4 forests (2 Breiman, 2 CRF) per layer (right), with respect to the maximal number of layers.

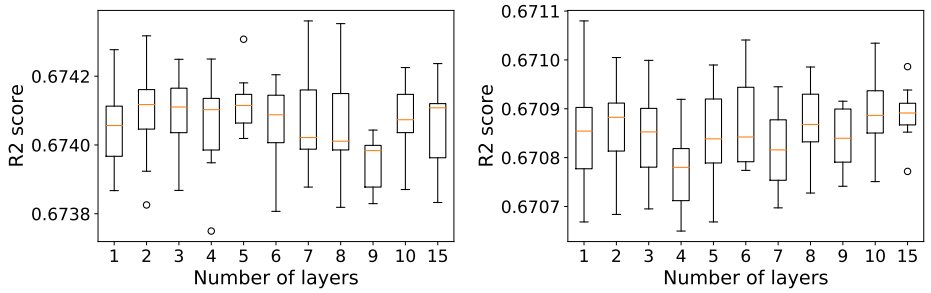



Figure S31: Airbnb dataset. Boxplots over 10 tries of the accuracy of a DF with 1 forest by layer (left) or 4 forests by layer (right), with respect to the DF maximal number of layers.

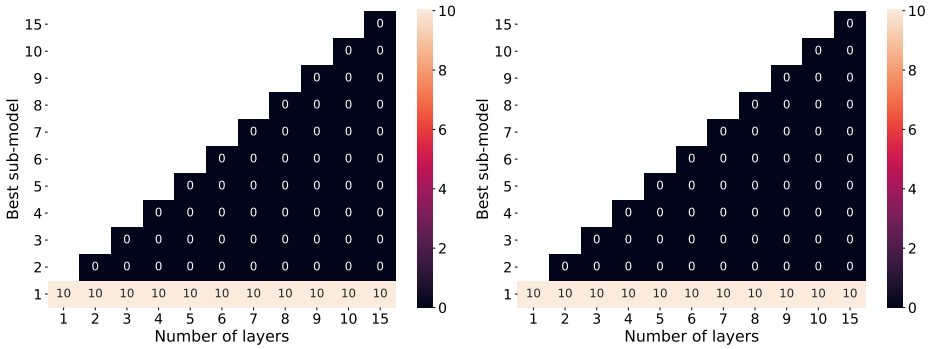



Figure S32: Airbnb datase. Heatmap counting the index of the sub-optimal model over 10 tries of a default DF with 1 (Breiman) forest per layer (left) or 4 forests (2 Breiman, 2 CRF) per layer (right), with respect to the maximal number of layers.

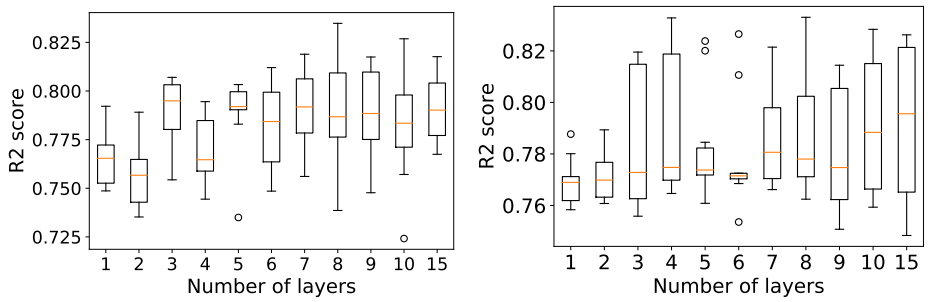



Figure S33: Housing dataset. Boxplots over 10 tries of the accuracy of a DF with 1 forest by layer (left) or 4 forests by layer (right), with respect to the DF maximal number of layers.

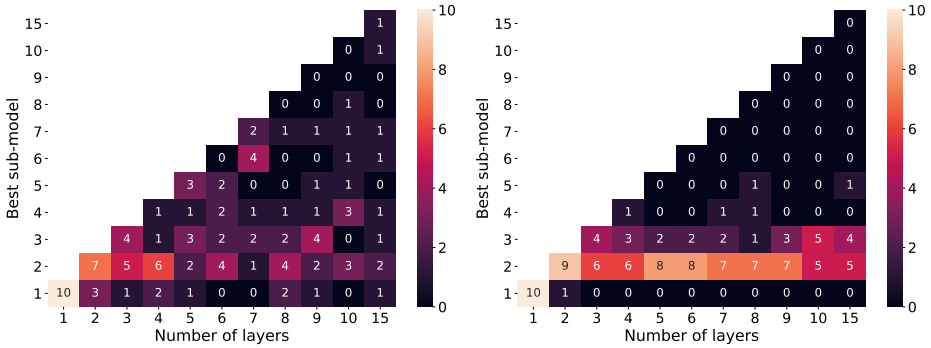



Figure S34: Housing dataset. Heatmap counting the index of the sub-optimal model over 10 tries of a default DF with 1 (Breiman) forest per layer (left) or 4 forests (2 Breiman, 2 CRF) per layer (right), with respect to the maximal number of layers.

## S1.4 Additional figures to Section 4

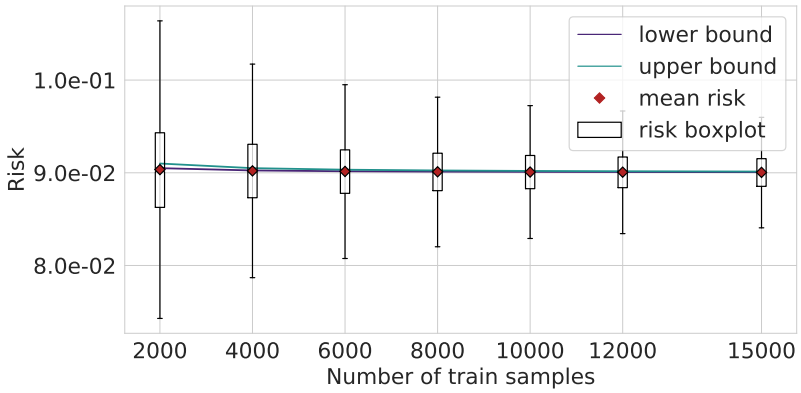



Figure S35: Illustration of the theoretical bounds of Proposition 2, when the first-layer tree is of depth $k = 2$ (when $k^\star = 4$) and $p = 0.8$. We draw a sample of size $n$ (x-axis), and a shallow tree network $r_{k,1,n}$ is fitted for which the theoretical risk is evaluated. Each boxplot is built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

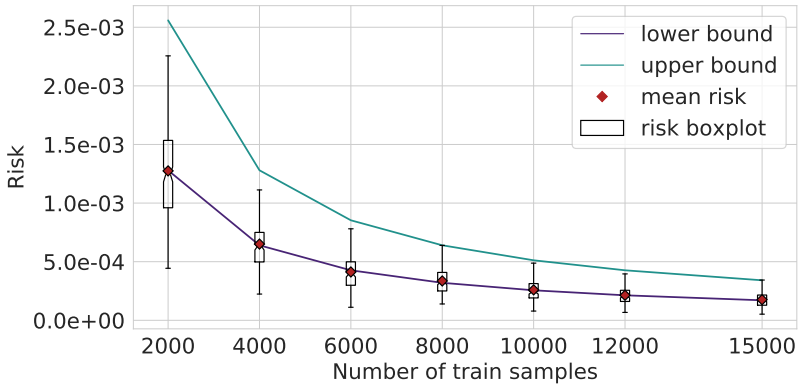



Figure S36: Illustration of the theoretical bounds of Proposition 2, when the first-layer tree is of depth $k = 4$ (when $k^\star = 4$) and $p = 0.8$. We draw a sample of size $n$ (x-axis), and a shallow tree network $r_{k,1,n}$ is fitted for which the theoretical risk is evaluated. Each boxplot is built out of 20 000 repetitions. The outliers are not shown for the sake of presentation.

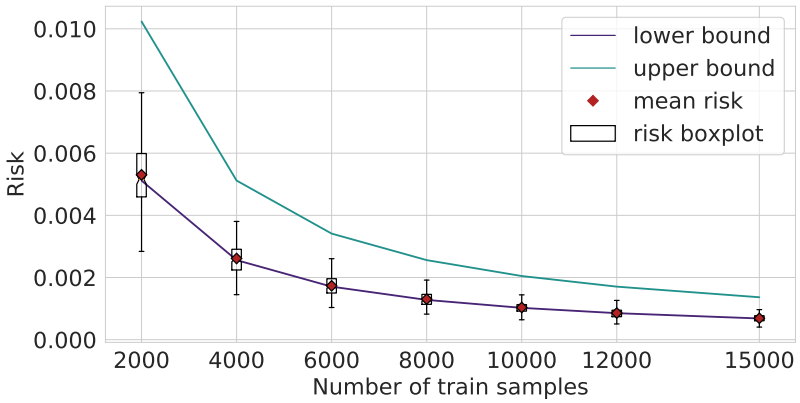



Figure S37: Illustration of the theoretical bounds of Proposition 2, when the first-layer tree is of depth $k = 6$ (when $k^\star = 4$) and $p = 0.8$. We draw a sample of size $n$ (x-axis), and a shallow tree network $r_{k,1,n}$ is fitted for which the theoretical risk is evaluated. Each boxplot is built out of 20 000 repetitions. The outliers are not shown for the sake of presentation.

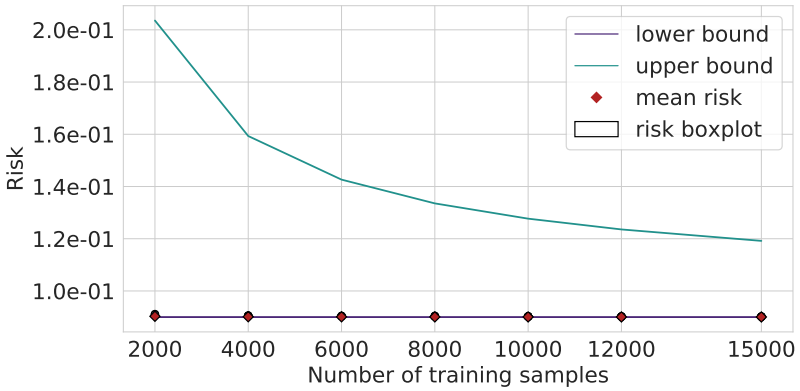



Figure S38: Illustration of the theoretical bounds of Proposition 3, when the first-layer tree is of depth $k = 2$ (when $k^\star = 4$) and $p = 0.8$. We draw a sample of size $n$ (x-axis), and a shallow tree network $r_{k,1,n}$ is fitted for which the theoretical risk is evaluated. Each boxplot is built 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.

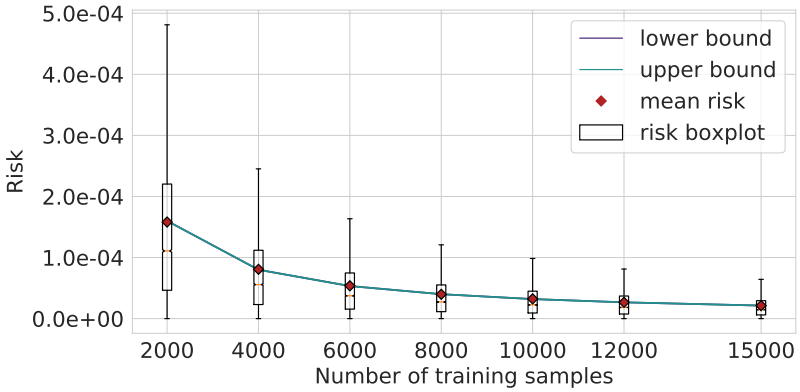



Figure S39: Illustration of the theoretical bounds of Proposition 3, when the first-layer tree is of depth $k = 4$ (when $k^\star = 4$) and $p = 0.8$. We draw a sample of size $n$ (x-axis), and a shallow tree network $r_{k,1,n}$ is fitted for which the theoretical risk is evaluated. Each boxplot is built out of 20 000 repetitions. The outliers are not shown for the sake of presentation. Note that the lower bound and the upper bound are merged.

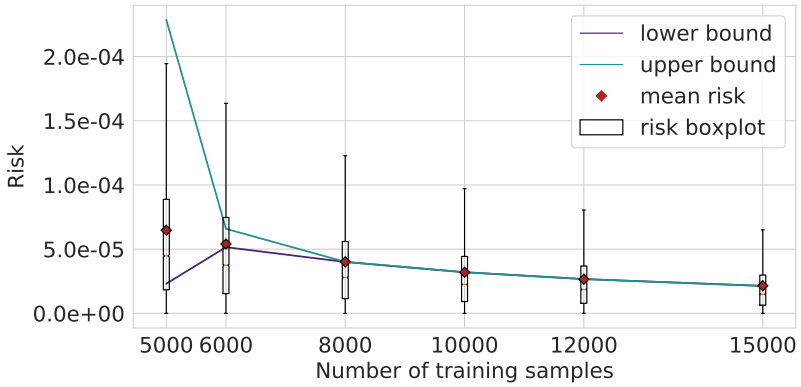



Figure S40: Illustration of the theoretical bounds of Proposition 3, when the first-layer tree is of depth $k = 6$ (when $k^\star = 4$) and $p = 0.8$. We draw a sample of size $n$ (x-axis), and a shallow tree network $r_{k,1,n}$ is fitted for which the theoretical risk is evaluated. Each boxplot is built out of 20 000 repetitions. The outliers are not shown for the sake of presentation.

## S2 Technical results on binomial random variables

**Lemma S4.** *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 $\leq n$.*
$$\mathbb{E}\left[Z \mid Z \geq k\right] = np + (1-p)k\frac{\mathbb{P}\left(Z=k\right)}{\sum_{i=k}^{n}\mathbb{P}\left(Z=i\right)}$$

*(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 [4, Lemma 11] 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 [7, Section 1.].

**(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 [7, Section 2.] 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 :

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

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 S5** (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)

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

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

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

Noticing that $N_n(C_n(X))$ is a binomial $\mathfrak{B}(n, \frac{1}{2^k})$, we obtain the upper bound using Lemma S4 (i) :

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

the lower bound is immediately obtained by applying Lemma S4, (i):

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

□

# S3 Proof of Lemma 1

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) When $k < k^\star$, the first tree is biased, since the optimal depth is $k^\star$. The second tree has access to the raw features or to the new feature created by the first tree. Since, for all leaves $C$ of the first tree, $\mathbb{P}[Y=1|X\in C]=0.5$, the new feature created by the first tree is non-informative (since it is constant, equal to 0.5). Therefore, the second-layer may use only raw feature and is consequently optimal if and only if $k' \geq k^\star$.

(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 this two regions leads to a shallow tree network of risk zero.

# S4 Proof of Proposition 2

1. Assume that $k < k^\star$. Recall that if a cell is empty, the tree prediction in this cell is set (arbitrarily) to zero. Thus,

$$\begin{aligned}
&\mathbb{E}\left[(r_{k,0,n}(X)-r(X))^2\right] \\
&= \mathbb{E}\left[(r_{k,0,n}(X)-r(X))^2\mathbb{1}_{N_n(C_n(X))>0}\right] + \mathbb{E}\left[(r(X))^2\mathbb{1}_{N_n(C_n(X))=0}\right], &(15)\\
&= \mathbb{E}\left[\left(\frac{1}{N_n(C_n(X))}\sum_{X_i\in C_n(X)} Y_i - r(X)\right)^2 \mathbb{1}_{N_n(C_n(X))>0}\right] + \mathbb{E}\left[(r(X))^2\mathbb{1}_{N_n(C_n(X))=0}\right], &(16)
\end{aligned}$$

where

$$\mathbb{E}\left[(r(X))^2 \mathbb{1}_{N_n(C_n(X))=0}\right] = \mathbb{E}\left[(r(X))^2 \mathbb{1}_{N_n(C_n(X))=0} \mathbb{1}_{X\in\mathcal{B}}\right] + \mathbb{E}\left[(r(X))^2 \mathbb{1}_{N_n(C_n(X))=0} \mathbb{1}_{X\in\mathcal{W}}\right] \tag{17}$$

$$= \left(\frac{p^2}{2} + \frac{(1-p)^2}{2}\right) \mathbb{P}\left(N_n(C_n(X)) = 0\right) \tag{18}$$

$$= (p^2 + (1-p)^2)\frac{(1-2^{-k})^n}{2}. \tag{19}$$

We now study the first term in (16), by considering that $X$ falls into $\mathcal{B}$ (the same computation holds when $X$ falls into $\mathcal{W}$). Letting $(X', Y')$ a generic random variable with the same distribution as $(X, Y)$, one has

$$\mathbb{E}\left[\left(\frac{1}{N_n(C_n(X))} \sum_{X_i \in C_n(X)} Y_i - p\right)^2 \mathbb{1}_{N_n(C_n(X))>0} \mathbb{1}_{X\in\mathcal{B}}\right] \tag{20}$$

$$= \frac{1}{2}\mathbb{E}\left[\left(\frac{1}{N_n(C_n(X))} \sum_{X_i \in C_n(X)} \left(Y_i - \mathbb{E}\left[Y' | X' \in C_n(X)\right]\right)\right)^2 \mathbb{1}_{N_n(C_n(X))>0}\right] \tag{21}$$

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

$$= \frac{1}{2}\mathbb{E}\left[\frac{\mathbb{1}_{N_n(C_n(X))>0}}{N_n(C_n(X))^2}\mathbb{E}\left[\left(\sum_{X_i \in C_n(X)} \left(Y_i - \mathbb{E}\left[Y' | X' \in C_n(X)\right]\right)\right)^2 \mid N_n(C_n(X))\right]\right]$$

$$+ \frac{1}{2}\left(p - \frac{1}{2}\right)^2 \mathbb{P}\left(N_n(C_n(X)) > 0\right), \tag{22}$$

where we used the fact that $\mathbb{E}\left[Y' | X' \in C_n(X)\right] = 1/2$ as in any leaf there is the same number of black and white cells. Moreover, conditional to $N_n(C_n(X))$, $\sum_{X_i \in C_n(X)} Y_i$ is a binomial random variable with parameters $\mathfrak{B}(N_n(C_n(X)), \frac{1}{2})$. Hence we obtain :

$$\mathbb{E}\left[\frac{\mathbb{1}_{N_n(C_n(X))>0}}{N_n(C_n(X))^2}\mathbb{E}\left[\left(\sum_{X_i \in C_n(X)} \left(Y_i - \mathbb{E}\left[Y' | X' \in C_n(X)\right]\right)\right)^2 | N_n(C_n(X))\right]\right] \tag{23}$$

$$= \frac{1}{4}\mathbb{E}\left[\frac{\mathbb{1}_{N_n(C_n(X))>0}}{N_n(C_n(X))}\right]. \tag{24}$$

The same computation holds when $X$ falls into $\mathcal{W}$. Indeed, the left-hand side term in (22) is unchanged, as for the right-hand side term, note that $(\frac{1}{2} - p)^2 = (\frac{1}{2} - (1-p))^2$. Consequently,

$$\mathbb{E}\left[\left(\frac{1}{N_n(C_n(X))} \sum_{X_i \in C_n(X)} Y_i - r(X)\right)^2 \mathbb{1}_{N_n(C_n(X))>0}\right] \tag{25}$$

$$= \frac{1}{4}\mathbb{E}\left[\frac{\mathbb{1}_{N_n(C_n(X))>0}}{N_n(C_n(X))}\right] + \left(p - \frac{1}{2}\right)^2 (1 - (1-2^{-k})^n). \tag{26}$$

Injecting (26) into (16), we have

$$\mathbb{E}\left[(r_{k,0,n}(X) - r(X))^2\right] \tag{27}$$

$$= \frac{1}{4}\mathbb{E}\left[\frac{\mathbb{1}_{N_n(C_n(X))>0}}{N_n(C_n(X))}\right] + \left(p - \frac{1}{2}\right)^2 (1 - (1-2^{-k})^n) + (p^2 + (1-p)^2)\frac{(1-2^{-k})^n}{2} \tag{28}$$

$$= \frac{1}{4}\mathbb{E}\left[\frac{\mathbb{1}_{N_n(C_n(X))>0}}{N_n(C_n(X))}\right] + \left(p - \frac{1}{2}\right)^2 + \left(p^2 + (1-p)^2 - 2\left(p - \frac{1}{2}\right)^2\right)\frac{(1-2^{-k})^n}{2} \tag{29}$$

$$= \frac{1}{4}\mathbb{E}\left[\frac{\mathbb{1}_{N_n(C_n(X))>0}}{N_n(C_n(X))}\right] + \left(p - \frac{1}{2}\right)^2 + \frac{(1-2^{-k})^n}{4}. \tag{30}$$

Noticing that $N_n(C_n(X))$ is a binomial random variable $\mathfrak{B}(n, \frac{1}{2^k})$, we obtain the upper and lower bounds with Lemma S4 (i):

$$\mathbb{E}\left[\frac{\mathbb{1}_{N_n(C_n(X))>0}}{N_n(C_n(X))}\right] \leq \frac{2^{k+1}}{n+1}, \tag{31}$$

and,

$$\mathbb{E}\left[\frac{\mathbb{1}_{N_n(C_n(X))>0}}{N_n(C_n(X))}\right] \geq \left(1 - (1-2^{-k})^n\right)\frac{2^k}{n+1}. \tag{32}$$

Gathering all the terms gives the result,

$$\mathbb{E}\left[(r_{k,0,n}(X) - r(X))^2\right] \leq \left(p - \frac{1}{2}\right)^2 + \frac{2^k}{2(n+1)} + \frac{(1-2^{-k})^n}{4}$$

and

$$\mathbb{E}\left[(r_{k,0,n}(X) - r(X))^2\right] \geq \left(p - \frac{1}{2}\right)^2 + \frac{2^k}{4(n+1)} + \frac{(1-2^{-k})^n}{4}\left(1 - \frac{2^k}{n+1}\right).$$

2. As in the proof of 1., we distinguish the case where the cell containing $X$ might be empty, in such a case the tree will predict 0:

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

$$= \mathbb{E}\left[(r_{k,0,n}(X) - r(X))^2 \mathbb{1}_{N_n(C_n(X))>0}\right] + \mathbb{E}\left[(r(X))^2 \mathbb{1}_{N_n(C_n(X))=0}\right] \tag{33}$$

$$= \mathbb{E}\left[(r_{k,0,n}(X) - r(X))^2 \mathbb{1}_{N_n(C_n(X))>0}\right] + (p^2 + (1-p)^2)\frac{(1-2^{-k})^n}{2}. \tag{34}$$

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

We have

$$\begin{aligned}
&\mathbb{E}\left[(r_{k,0,n}(X)-p)^2)\mathbb{1}_{X\in\mathcal{B}}\mathbb{1}_{N_n(C_n(X))>0}\right] \\
&= \sum_{L_j\subset\mathcal{B}} \mathbb{E}\left[\left(\frac{\mathbb{1}_{N_n(L_j)>0}}{N_n(L_j)}\sum_{X_i\in L_j}(Y_i-p)\right)^2 \mathbb{1}_{X\in L_j}\right] && (35)\\
&= \frac{2^k}{2}\cdot\mathbb{E}\left[\left(\frac{\mathbb{1}_{N_n(L_b)>0}}{N_n(L_b)}\sum_{X_i\in L_b}(Y_i-p)\right)^2\right]\mathbb{P}\left(X\in L_b\right) && (36)\\
&= \frac{1}{2}\mathbb{E}\left[\left(\frac{\mathbb{1}_{N_n(L_b)>0}}{N_n(L_b)}\sum_{X_i\in L_b}(Y_i-p)\right)^2\right] && (37)\\
&= \frac{1}{2}\mathbb{E}\left[\frac{\mathbb{1}_{N_n(L_b)>0}}{N_n(L_b)^2}\mathbb{E}\left[\left(\sum_{X_i\in L_b}(Y_i-p)\right)^2|N_n(L_b)\right]\right] && (38)\\
&= \frac{1}{2}\mathbb{E}\left[\frac{\mathbb{1}_{N_n(L_b)>0}}{N_n(L_b)^2}\mathbb{E}\left[\sum_{X_i\in L_b}(Y_i-p)^2|N_n(L_b)\right]\right] \qquad \text{(by independence of the } Y_i\text{)} && (39)\\
&= \frac{1}{2}\mathbb{E}\left[\frac{\mathbb{1}_{N_n(L_b)>0}}{N_n(L_b)}p(1-p)\right]. && (40)
\end{aligned}$$

Remark that the above computation holds when $X\in\mathcal{W}$ after replacing $p$ by $(1-p)$, $B$ by $W$ 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 (34), the computation in (40) 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, as in the proof of 1., we can conclude using Lemma S4 (i) :

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

and

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

# S5 Proof of Proposition 3

Let $k\in\mathbb{N}$. Denote by $\mathcal{L}_k=\{L_{i,k},i=1,\dots,2^k\}$ 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

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

## S5.1 Proof of 1. (i) (lower-bound for the case $k < k^\star$)

Recall that $k < k^\star$. In this case, each leaf of the encoding tree is contains half black square and half white square (see Figure 8a). Hence, the empirical probability of $Y$ being equal to one in such leaf is close to $1/2$. Recalling that our estimate is $r_{k,1,n}$, we have

$$\begin{aligned}
&\mathbb{E}\left[(r_{k,1,n}(X) - r(X))^2\right] \\
&= \mathbb{E}\left[(r_{k,1,n}(X) - p)^2 \mathbb{1}_{X \in \mathcal{B}} \mathbb{1}_{X \in \tilde{\mathcal{B}}_k}\right] + \mathbb{E}\left[(r_{k,1,n}(X) - p)^2 \mathbb{1}_{X \in \mathcal{B}} \mathbb{1}_{X \in \tilde{\mathcal{W}}_k}\right] \\
&+ \mathbb{E}\left[(r_{k,1,n}(X) - (1-p))^2 \mathbb{1}_{X \in \mathcal{W}} \mathbb{1}_{X \in \tilde{\mathcal{B}}_k}\right] + \mathbb{E}\left[(r_{k,1,n}(X) - (1-p))^2 \mathbb{1}_{X \in \mathcal{W}} \mathbb{1}_{X \in \tilde{\mathcal{W}}_k}\right] \qquad (41) \\
&+ \mathbb{E}\left[(r_{k,1,n}(X) - p)^2 \mathbb{1}_{X \in \mathcal{B}} (1 - \mathbb{1}_{X \in \tilde{\mathcal{B}}_k} - \mathbb{1}_{X \in \tilde{\mathcal{W}}_k})\right] + \mathbb{E}\left[(r_{k,1,n}(X) - (1-p))^2 \mathbb{1}_{X \in \mathcal{W}} (1 - \mathbb{1}_{X \in \tilde{\mathcal{B}}_k} - \mathbb{1}_{X \in \tilde{\mathcal{W}}_k})\right]
\end{aligned}$$

Note that $X \notin \tilde{\mathcal{B}}_k \cup \tilde{\mathcal{W}}_k$ is equivalent to $X$ belonging to an empty cell. Besides, the prediction is null by convention in an empty cell. Therefore, the sum of the last two terms in (41) can be written as

$$\mathbb{E}\left[p^2 \mathbb{1}_{X \in \mathcal{B}} \mathbb{1}_{N_n(C_n(X))=0})\right] + \mathbb{E}\left[(1-p)^2 \mathbb{1}_{X \in \mathcal{W}} \mathbb{1}_{N_n(C_n(X))=0})\right] = \frac{p^2 + (1-p)^2}{2} \left(1 - \frac{1}{2^k}\right)^n. \qquad (42)$$

To begin with we focus on the first two terms in (41). We deal with the last two terms at the very end as similar computations are conducted.

$$\begin{aligned}
&\mathbb{E}\left[(r_{k,1,n}(X) - p)^2 \mathbb{1}_{X \in \mathcal{B}} \mathbb{1}_{X \in \tilde{\mathcal{B}}_k}\right] + \mathbb{E}\left[(r_{k,1,n}(X) - p)^2 \mathbb{1}_{X \in \mathcal{B}} \mathbb{1}_{X \in \tilde{\mathcal{W}}_k}\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\left(\frac{1}{N_n(\tilde{\mathcal{B}}_k)} \sum_{X_i \in \tilde{\mathcal{B}}_k} Y_i - p\right)^2 \middle| \tilde{\mathcal{B}}_k\right] \mathbb{P}\left(X \in \tilde{\mathcal{B}}_k, X \in \mathcal{B} | \tilde{\mathcal{B}}_k\right)\right] \\
&\quad + \mathbb{E}\left[\mathbb{E}\left[\left(\frac{1}{N_n(\tilde{\mathcal{W}}_k)} \sum_{X_i \in \tilde{\mathcal{W}}_k} Y_i - p\right)^2 \middle| \tilde{\mathcal{W}}_k\right] \mathbb{P}\left(X \in \tilde{\mathcal{W}}_k, X \in \mathcal{B} | \tilde{\mathcal{W}}_k\right)\right]. \qquad (43)
\end{aligned}$$

Regarding the left-hand side term in (43),

$$\mathbb{E}\left[\left(\frac{1}{N_n(\tilde{\mathcal{B}}_k)} \sum_{X_i \in \tilde{\mathcal{B}}_k} Y_i - p\right)^2 \middle| \tilde{\mathcal{B}}_k\right] \leq \left(p - \frac{1}{2}\right)^2, \qquad (44)$$

since $p > 1/2$ and, by definition of $\tilde{\mathcal{B}}_k$,

$$\sum_{X_i \in \tilde{\mathcal{B}}_k} Y_i \geq N_n(\tilde{\mathcal{B}}_k)/2.$$

Now, regarding right-hand side term in (43), we let

$$Z_{\tilde{\mathcal{W}}_k} = \mathbb{E}\left[\sum_{X_i \in \tilde{\mathcal{W}}_k} Y_i \mid N_1, ..., N_{2^k}, \tilde{\mathcal{W}}_k\right],$$

where $N_1, ..., N_{2^k}$ denote the number of data points falling in each leaf $L_1, \ldots, L_{2^k}$ of the encoding tree. Hence,

$$\mathbb{E}\left[\left(\frac{1}{N_n(\tilde{\mathcal{W}}_k)}\sum_{X_i\in\tilde{\mathcal{W}}_k} Y_i - p\right)^2 \Big| \tilde{\mathcal{W}}_k\right] = \mathbb{E}\left[\frac{1}{N_n(\tilde{\mathcal{W}}_k)^2}\mathbb{E}\left[\left(\sum_{X_i\in\tilde{\mathcal{W}}_k} Y_i - Z_{\tilde{\mathcal{W}}_k}\right)^2 + \left(Z_{\tilde{\mathcal{W}}_k} - N_n(\tilde{\mathcal{W}}_k)p\right)^2 \right.\right.$$
$$\left.\left. +2\left(\sum_{X_i\in\tilde{\mathcal{W}}_k} Y_i - Z_{\tilde{\mathcal{W}}_k}\right)\left(Z_{\tilde{\mathcal{W}}_k} - N_n(\tilde{\mathcal{W}}_k)p\right) \mid N_1, ..., N_{2^k}, \tilde{\mathcal{W}}_k\right] \Big| \tilde{\mathcal{W}}_k\right] \tag{45}$$

The cross-term is null according to the definition of $Z_{\tilde{\mathcal{W}}_k}$, and since $(Z_{\tilde{\mathcal{W}}_k} - N_n(\tilde{\mathcal{W}}_k))$ is $(N_1, ..., N_{2^k}, \tilde{\mathcal{W}}_k)$-measurable. Therefore,

$$\begin{aligned}\mathbb{E}\left[\left(\frac{1}{N_n(\tilde{\mathcal{W}}_k)}\sum_{X_i\in\tilde{\mathcal{W}}_k} Y_i - p\right)^2 \Big| \tilde{\mathcal{W}}_k\right] &= \mathbb{E}\left[\frac{1}{N_n(\tilde{\mathcal{W}}_k)^2}\mathbb{E}\left[\left(\sum_{X_i\in\tilde{\mathcal{W}}_k} Y_i - Z_{\tilde{\mathcal{W}}_k}\right)^2 \mid N_1, ..., N_{2^k}, \tilde{\mathcal{W}}_k\right] \Big| \tilde{\mathcal{W}}_k\right] \\ &\quad + \mathbb{E}\left[\frac{1}{N_n(\tilde{\mathcal{W}}_k)^2}\mathbb{E}\left[\left(Z_{\tilde{\mathcal{W}}_k} - N_n(\tilde{\mathcal{W}}_k)p\right)^2 \mid N_1, ..., N_{2^k}, \tilde{\mathcal{W}}_k\right] \Big| \tilde{\mathcal{W}}_k\right] \\ &= I_n + J_n, \end{aligned} \tag{46}$$

where $I_n$ and $J_n$ can be respectively identified as variance and bias terms. Indeed,

$$\mathbb{E}\left[\left(\sum_{X_i\in\tilde{\mathcal{W}}_k} Y_i - Z_{\tilde{\mathcal{W}}_k}\right)^2 \mid N_1, ..., N_{2^k}, \tilde{\mathcal{W}}_k\right]$$

is the variance of a binomial random variable $B(N_n(\tilde{\mathcal{W}}_k), \frac{1}{2})$ conditioned to be lower or equal to $N_n(\tilde{\mathcal{W}}_k)/2$. According to Technical Lemma S6, we have

$$I_n \leq \frac{1}{4}\mathbb{E}\left[\frac{\mathbb{1}_{N_n(\tilde{\mathcal{W}}_k)>0}}{N_n(\tilde{\mathcal{W}}_k)\mathbb{P}\left(B(N_n(\tilde{\mathcal{W}}_k), 1/2) \leq N_n(\tilde{\mathcal{W}}_k)/2\right)} \Big| \tilde{\mathcal{W}}_k\right] \leq \frac{1}{2}\mathbb{E}\left[\frac{\mathbb{1}_{N_n(\tilde{\mathcal{W}}_k)>0}}{N_n(\tilde{\mathcal{W}}_k)} \Big| \tilde{\mathcal{W}}_k\right]. \tag{47}$$

Regarding $J_n$,

$$Z_{\tilde{\mathcal{W}}_k} - N_n(\tilde{\mathcal{W}}_k)p = \mathbb{E}\left[\sum_{X_i\in\tilde{\mathcal{W}}_k} Y_i \mid N_1, ..., N_{2^k}, \tilde{\mathcal{W}}_k\right] - N_n(\tilde{\mathcal{W}}_k)p \tag{48}$$

$$= \mathbb{E}\left[\sum_{j=1}^{2^k}\sum_{X_i\in L_j} Y_i \mathbb{1}_{L_j\subset\tilde{\mathcal{W}}_k} \mid N_1, ..., N_{2^k}, \tilde{\mathcal{W}}_k\right] - N_n(\tilde{\mathcal{W}}_k)p \tag{49}$$

$$= \sum_{j=1}^{2^k}\left(\mathbb{E}\left[\sum_{X_i\in L_j} Y_i \mid N_1, ..., N_{2^k}, \tilde{\mathcal{W}}_k\right] - pN_j\right)\mathbb{1}_{L_j\subset\tilde{\mathcal{W}}_k}, \tag{50}$$

since $\mathbb{1}_{L_j\subset\tilde{\mathcal{W}}_k}$ is $\tilde{\mathcal{W}}_k$ -measurable and $N_n(\tilde{\mathcal{W}}_k) = \sum_{i=1}^{2^k} N_j$. Noticing that

$$\mathbb{E}\left[\sum_{X_i\in L_j} Y_i \mid N_1, ..., N_{2^k}, \tilde{\mathcal{W}}_k\right] = \mathbb{E}\left[\sum_{X_i\in L_j} Y_i \mid N_j, \tilde{\mathcal{W}}_k\right], \tag{51}$$

we deduce

$$Z_{\tilde{\mathcal{W}}_k} - N_n(\tilde{\mathcal{W}}_k)p = \sum_{j=1}^{2^k}\left(\mathbb{E}\left[\sum_{X_i\in L_j} Y_i \mid N_j, \tilde{\mathcal{W}}_k\right] - N_j p\right)\mathbb{1}_{L_j\subset\tilde{\mathcal{W}}_k} \tag{52}$$

and

$$(Z_{\tilde{\mathcal{W}}_k} - N_n(\tilde{\mathcal{W}}_k)p)^2 = \left(\sum_{j=1}^{2^k} f_j \mathbb{1}_{L_j \subset \tilde{\mathcal{W}}_k}\right)^2 \tag{53}$$

with $f_j = \left(N_j p - \mathbb{E}\left[\sum_{X_i \in L_j} Y_i \mid N_j, \tilde{\mathcal{W}}_k\right]\right)$. For all $j$, such that $L_j \subset \tilde{\mathcal{W}}_k$, $\mathbb{E}\left[\sum_{X_i \in L_j} Y_i \mid N_j, \tilde{\mathcal{W}}_k\right]$ is a binomial random variable $\mathfrak{B}(N_n(\tilde{\mathcal{W}}_k), \frac{1}{2})$ conditioned to be lower or equal to $N_n(\tilde{\mathcal{W}}_k)/2$. Using Lemma S4 (vi), we obtain :

$$f_j \leq N_j\left(p - \frac{1}{2}\right) + \sqrt{N_j}\left(\frac{1}{\sqrt{\pi}} + \frac{2\sqrt{2}}{\pi\sqrt{(2n+1)}}\right) \tag{54}$$

$$\leq N_j\left(p - \frac{1}{2}\right) + \sqrt{N_j} + \frac{2}{\pi}. \tag{55}$$

Therefore,

$$(Z_{\tilde{\mathcal{W}}_k} - N_n(\tilde{\mathcal{W}}_k)p)^2 \leq \left(N_n(\tilde{\mathcal{W}}_k)\left(p - \frac{1}{2}\right) + \sum_{j=1}^{2^k}\sqrt{N_j}\mathbb{1}_{L_j \subset \tilde{\mathcal{W}}_k} + \frac{2^{k+1}}{\pi}\right)^2 \tag{56}$$

$$\leq \left(N_n(\tilde{\mathcal{W}}_k)\left(p - \frac{1}{2}\right) + 2^{k/2}\sqrt{N_n(\tilde{\mathcal{W}}_k)} + \frac{2^{k+1}}{\pi}\right)^2, \tag{57}$$

since, according to Cauchy-Schwarz inequality,

$$\sum_{j=1}^{2^k}\sqrt{N_j}\mathbb{1}_{L_j \subset \tilde{\mathcal{W}}_k} \leq 2^{k/2} N_n(\tilde{\mathcal{W}}_k)^{1/2}. \tag{58}$$

Overall

$$J_n \leq \mathbb{E}\left[\frac{1}{N_n(\tilde{\mathcal{W}}_k)^2}\mathbb{E}\left[\left(N_n(\tilde{\mathcal{W}}_k)\left(p - \frac{1}{2}\right) + 2^{k/2}N_n(\tilde{\mathcal{W}}_k)^{1/2} + \frac{2^{k+1}}{\pi}\right)^2 \mid N_1, ..., N_{2^k}, \tilde{\mathcal{W}}_k\right]\Big|\tilde{\mathcal{W}}_k\right] \tag{59}$$

$$\leq \left(p - \frac{1}{2}\right)^2 + 2^k\mathbb{E}\left[\frac{\mathbb{1}_{N_n(\tilde{\mathcal{W}}_k)>0}}{N_n(\tilde{\mathcal{W}}_k)}\Big|\tilde{\mathcal{W}}_k\right] + \frac{2^{2k+2}}{\pi^2}\mathbb{E}\left[\frac{\mathbb{1}_{N_n(\tilde{\mathcal{W}}_k)>0}}{N_n(\tilde{\mathcal{W}}_k)^2}\Big|\tilde{\mathcal{W}}_k\right] + 2^{k/2+1}\left(p - \frac{1}{2}\right)\mathbb{E}\left[\frac{\mathbb{1}_{N_n(\tilde{\mathcal{W}}_k)>0}}{N_n(\tilde{\mathcal{W}}_k)^{1/2}}\Big|\tilde{\mathcal{W}}_k\right] \tag{60}$$

$$+ \frac{2^{k+2}}{\pi}\left(p - \frac{1}{2}\right)\mathbb{E}\left[\frac{\mathbb{1}_{N_n(\tilde{\mathcal{W}}_k)>0}}{N_n(\tilde{\mathcal{W}}_k)}\Big|\tilde{\mathcal{W}}_k\right] + \frac{2^{\frac{3k}{2}+2}}{\pi}\mathbb{E}\left[\frac{\mathbb{1}_{N_n(\tilde{\mathcal{W}}_k)>0}}{N_n(\tilde{\mathcal{W}}_k)^{3/2}}\Big|\tilde{\mathcal{W}}_k\right]. \tag{61}$$

All together, we obtain

$$I_n + J_n \leq \left(p - \frac{1}{2}\right)^2 + \left(2^k + \frac{1}{2} + \frac{2^{k+2}}{\pi}\left(p - \frac{1}{2}\right)\right)\mathbb{E}\left[\frac{\mathbb{1}_{N_n(\tilde{\mathcal{W}}_k)>0}}{N_n(\tilde{\mathcal{W}}_k)}\Big|\tilde{\mathcal{W}}_k\right] + \frac{2^{2k+2}}{\pi^2}\mathbb{E}\left[\frac{\mathbb{1}_{N_n(\tilde{\mathcal{W}}_k)>0}}{N_n(\tilde{\mathcal{W}}_k)^2}\Big|\tilde{\mathcal{W}}_k\right]$$
$$+ 2^{k/2+1}\left(p - \frac{1}{2}\right)\mathbb{E}\left[\frac{\mathbb{1}_{N_n(\tilde{\mathcal{W}}_k)>0}}{N_n(\tilde{\mathcal{W}}_k)^{1/2}}\Big|\tilde{\mathcal{W}}_k\right] + \frac{2^{\frac{3k}{2}+2}}{\pi}\mathbb{E}\left[\frac{\mathbb{1}_{N_n(\tilde{\mathcal{W}}_k)>0}}{N_n(\tilde{\mathcal{W}}_k)^{3/2}}\Big|\tilde{\mathcal{W}}_k\right]$$

We apply Lemma S4(i)(iv) to $N_n(\tilde{\mathcal{W}}_k)$ which is a binomial $\mathfrak{B}(n, p')$ where $p' = \mathbb{P}(X \in \tilde{\mathcal{W}}_k | \tilde{\mathcal{W}}_k)$ :

$$\mathbb{E}\left[\frac{\mathbb{1}_{N_n(\tilde{\mathcal{W}}_k)>0}}{N_n(\tilde{\mathcal{W}}_k)}\Big|\tilde{\mathcal{W}}_k\right] \leq \frac{2}{(n+1)p'},$$

$$\mathbb{E}\left[\frac{\mathbb{1}_{N_n(\tilde{\mathcal{W}}_k)>0}}{N_n(\tilde{\mathcal{W}}_k)^{1/2}}\Big|\tilde{\mathcal{W}}_k\right] \leq \frac{2}{\sqrt{n \cdot p'}}.$$

We deduce that

$$I_n + J_n \leq (p-\frac{1}{2})^2 + \frac{2^{k/2+2}(p-\frac{1}{2})}{\sqrt{\pi n \cdot p'}} + \frac{2}{(n+1)\cdot p'}\left(2^k + \frac{1}{2} + \frac{2^{k+2}}{\pi} + \frac{2^{3k/2+2}}{\pi\sqrt{\pi}} + 3\cdot\frac{2^{2k+2}}{\pi^2}\right).$$

Finally,

$$\begin{aligned}
&\mathbb{E}\left[(r_{k,1,n}(X)-p)^2 \mathbb{1}_{X\in\mathcal{B}}\mathbb{1}_{X\in\tilde{\mathcal{B}}_k}\right] + \mathbb{E}\left[(r_{k,1,n}(X)-p)^2 \mathbb{1}_{X\in\mathcal{B}}\mathbb{1}_{X\in\tilde{\mathcal{W}}_k}\right] \\
&\leq \left(p-\frac{1}{2}\right)^2 \mathbb{P}\left(X\in\tilde{\mathcal{B}}_k, X\in\mathcal{B}\right) + \mathbb{E}\left[(I_n+J_n)\mathbb{P}\left(X\in\tilde{\mathcal{W}}_k, X\in\mathcal{B}|\tilde{\mathcal{W}}_k\right)\right]
\end{aligned}$$

Since for all $\tilde{\mathcal{B}}_k$, there is exactly the same number of black cells and white cells in $\tilde{\mathcal{B}}_k$, we have

$$\mathbb{P}\left(X\in\tilde{\mathcal{W}}_k, X\in\mathcal{B}|\tilde{\mathcal{W}}_k\right) = \frac{\mathbb{P}\left(X\in\tilde{\mathcal{W}}_k|\tilde{\mathcal{W}}_k\right)}{2},$$

yielding

$$\begin{aligned}
&\mathbb{E}\left[(r_{k,1,n}(X)-p)^2 \mathbb{1}_{X\in\mathcal{B}}\mathbb{1}_{X\in\tilde{\mathcal{B}}_k}\right] + \mathbb{E}\left[(r_{k,1,n}(X)-p)^2 \mathbb{1}_{X\in\mathcal{B}}\mathbb{1}_{X\in\tilde{\mathcal{W}}_k}\right] && (62)\\
\leq &\frac{1}{2}\left(p-\frac{1}{2}\right)^2 + \frac{2^{k/2+1}(p-\frac{1}{2})}{\sqrt{\pi n}} + \frac{1}{(n+1)}\left(2^k + \frac{1}{2} + \frac{2^{k+2}}{\pi} + \frac{2^{3k/2+2}}{\pi\sqrt{\pi}} + 3\cdot\frac{2^{2k+2}}{\pi^2}\right) && (63)\\
\leq &\frac{1}{2}\left(p-\frac{1}{2}\right)^2 + \frac{2^{k/2+1}(p-\frac{1}{2})}{\sqrt{\pi n}} + \frac{3\cdot 2^{2k+2}}{(n+1)\pi^2}(1+\varepsilon_1(k)) && (64)
\end{aligned}$$

where $\varepsilon_1(k) = \frac{\pi^2}{3\cdot 2^{(2k+2)}}\left(2^k + \frac{1}{2} + \frac{2^{k+2}}{\pi} + \frac{2^{3k/2+2}}{\pi\sqrt{\pi}}\right)$.

The two intermediate terms of (41) can be similarly bounded from above. Indeed,

$$\begin{aligned}
&\mathbb{E}\left[(r_{k,1,n}(X)-(1-p))^2 \mathbb{1}_{X\in\mathcal{W}}\mathbb{1}_{X\in\tilde{\mathcal{B}}_k}\right] + \mathbb{E}\left[(r_{k,1,n}(X)-(1-p))^2 \mathbb{1}_{X\in\mathcal{W}}\mathbb{1}_{X\in\tilde{\mathcal{W}}_k}\right] && (65)\\
&= \mathbb{E}\left[\mathbb{E}\left[\left(\frac{1}{N_n(\tilde{\mathcal{B}}_k)}\sum_{X_i\in\tilde{\mathcal{B}}_k} Y_i - (1-p)\right)^2 \middle| \tilde{\mathcal{B}}_k\right]\mathbb{P}\left(X\in\tilde{\mathcal{B}}_k, X\in\mathcal{W}|\tilde{\mathcal{B}}_k\right)\right] \\
&\quad + \mathbb{E}\left[\mathbb{E}\left[\left(\frac{1}{N_n(\tilde{\mathcal{W}}_k)}\sum_{X_i\in\tilde{\mathcal{W}}_k} Y_i - (1-p)\right)^2 \middle| \tilde{\mathcal{W}}_k\right]\mathbb{P}\left(X\in\tilde{\mathcal{W}}_k, X\in\mathcal{W}|\tilde{\mathcal{W}}_k\right)\right], && (66)
\end{aligned}$$

where, by definition of $\tilde{\mathcal{W}}_k$,

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

The first term in (66) can be treated similarly as above:

$$\begin{aligned}
\mathbb{E}\left[\left(\frac{1}{N_n(\tilde{\mathcal{B}}_k)}\sum_{X_i\in\tilde{\mathcal{B}}_k} Y_i - (1-p)\right)^2 \middle| \tilde{\mathcal{B}}_k\right] &= \mathbb{E}\left[\frac{1}{N_n(\tilde{\mathcal{B}}_k)^2}\mathbb{E}\left[\left(\sum_{X_i\in\tilde{\mathcal{B}}_k} Y_i - Z_{\tilde{\mathcal{B}}_k}\right)^2 \mid N_1,...,N_{2^k},\tilde{\mathcal{B}}_k\right]\middle|\tilde{\mathcal{B}}_k\right] \\
&\quad + \mathbb{E}\left[\frac{1}{N_n(\tilde{\mathcal{B}}_k)^2}\mathbb{E}\left[\left(Z_{\tilde{\mathcal{B}}_k} - N_n(\tilde{\mathcal{B}}_k)(1-p)\right)^2 \mid N_1,...,N_{2^k},\tilde{\mathcal{B}}_k\right]\middle|\tilde{\mathcal{B}}_k\right] \\
&= I_n' + J_n', && (67)
\end{aligned}$$

where

$$Z_{\tilde{\mathcal{B}}_k} = \mathbb{E}\left[\sum_{X_i\in\tilde{\mathcal{B}}_k} Y_i \mid N_1,...,N_{2^k},\tilde{\mathcal{B}}_k\right],$$

and the cross-term in (67) is null according to the definition of $Z_{\tilde{\mathcal{B}}_k}$. Regarding $I'_n$, note that

$$\mathbb{E}\left[\left(\sum_{X_i\in\tilde{\mathcal{B}}_k} Y_i - Z_{\tilde{\mathcal{B}}_k}\right)^2 \mid N_1,...,N_{2^k},\tilde{\mathcal{B}}_k\right]$$

is the variance of a binomial random variable $B(N_n(\tilde{\mathcal{B}}_k),\frac{1}{2})$ conditioned to be strictly larger than $N_n(\tilde{\mathcal{B}}_k)/2$. According to Technical Lemma S6, we have

$$I'_n \leq \frac{1}{4}\mathbb{E}\left[\frac{\mathbb{1}_{N_n(\tilde{\mathcal{B}}_k)>0}}{N_n(\tilde{\mathcal{B}}_k)\mathbb{P}\left(B(N_n(\tilde{\mathcal{B}}_k),1/2)>N_n(\tilde{\mathcal{B}}_k)/2\right)}\Big|\tilde{\mathcal{B}}_k\right] \leq \mathbb{E}\left[\frac{\mathbb{1}_{N_n(\tilde{\mathcal{B}}_k)>0}}{N_n(\tilde{\mathcal{B}}_k)}\Big|\tilde{\mathcal{B}}_k\right]. \tag{68}$$

To obtain the last inequality, notice that

$$\mathbb{P}\left(B(N_n(\tilde{\mathcal{B}}_k),1/2)>N_n(\tilde{\mathcal{B}}_k)/2\right) = \frac{1}{2}-\frac{1}{2}\mathbb{P}\left(B(N_n(\tilde{\mathcal{B}}_k),1/2)=N_n(\tilde{\mathcal{B}}_k)/2\right) \geq \frac{1}{2}\left(1-\frac{1}{\sqrt{\pi(n/2+1/4)}}\right) \geq \frac{1}{4}$$

as soon as $n\geq 4$.

Regarding $J'_n$, we have

$$\mathbb{E}\left[\frac{1}{N_n(\tilde{\mathcal{B}}_k)^2}\mathbb{E}\left[\left(Z_{\tilde{\mathcal{B}}_k}-N_n(\tilde{\mathcal{B}}_k)(1-p)\right)^2 \mid N_1,...,N_{2^k},\tilde{\mathcal{B}}_k\right]\right] \tag{69}$$

$$= \mathbb{E}\left[\frac{1}{N_n(\tilde{\mathcal{B}}_k)^2}\mathbb{E}\left[\left(\sum_{i=1}^{2^k}\left(\mathbb{E}\left[\sum_{X_i\in L_i} Y_i \mid N_j,\tilde{\mathcal{B}}_k\right]-N_j(1-p)\right)\mathbb{1}_{L_j\subset\tilde{\mathcal{B}}_k}\right)^2 \mid N_1,...,N_{2^k},\tilde{\mathcal{B}}_k\right]\right]. \tag{70}$$

For all $j$, such that $L_j\subset\tilde{\mathcal{B}}_k$, $\mathbb{E}\left[\sum_{X_i\in L_j} Y_i \mid N_j,\tilde{\mathcal{B}}_k\right]$ is a binomial random variable $\mathfrak{B}(N_j,\frac{1}{2})$ conditioned to be larger than $\lfloor (N_j+1)/2\rfloor$. Then, according to Technical Lemma (vii)

$$\mathbb{E}\left[\sum_{X_i\in L_j} Y_i \mid N_j,\tilde{\mathcal{B}}_k\right] \leq \frac{N_j}{2}+1+\frac{1}{\sqrt{\pi(N_j+1)}}.$$

Hence,

$$\mathbb{E}\left[\sum_{X_i\in L_i} Y_i \mid N_j,\tilde{\mathcal{B}}_k\right]-N_j(1-p) \leq N_j(p-\frac{1}{2})+1+\frac{1}{\sqrt{\pi(N_j+1)}} \tag{71}$$

$$\leq N_j\left(p-\frac{1}{2}\right)+\sqrt{N_j}+\frac{2}{\pi}, \tag{72}$$

for $N_j\geq 1$. Thus,

$$\mathbb{E}\left[\frac{1}{N_n(\tilde{\mathcal{B}}_k)^2}\mathbb{E}\left[\left(Z_{\tilde{\mathcal{B}}_k}-N_n(\tilde{\mathcal{B}}_k)(1-p)\right)^2 \mid N_1,...,N_{2^k},\tilde{\mathcal{B}}_k\right]\right] \tag{73}$$

$$\leq \mathbb{E}\left[\frac{1}{N_n(\tilde{\mathcal{B}}_k)^2}\mathbb{E}\left[\left(\sum_{i=1}^{2^k}\left(N_j\left(p-\frac{1}{2}\right)+\sqrt{N_j}+\frac{2}{\pi}\right)\mathbb{1}_{L_j\subset\tilde{\mathcal{B}}_k}\right)^2 \mid N_1,...,N_{2^k},\tilde{\mathcal{B}}_k\right]\right] \tag{74}$$

$$\leq \mathbb{E}\left[\frac{1}{N_n(\tilde{\mathcal{B}}_k)^2}\mathbb{E}\left[\left(N_n(\tilde{\mathcal{B}}_k)\left(p-\frac{1}{2}\right)+2^{k/2}\sqrt{N_n(\tilde{\mathcal{B}}_k)}+\frac{2^{k+1}}{\pi}\right)^2 \mid N_1,...,N_{2^k},\tilde{\mathcal{B}}_k\right]\right]. \tag{75}$$

All together, we obtain

$$I'_n+J'_n \leq \left(p-\frac{1}{2}\right)^2+\left(2^k+1+\frac{2^{k+2}}{\pi}\left(p-\frac{1}{2}\right)\right)\mathbb{E}\left[\frac{\mathbb{1}_{N_n(\tilde{\mathcal{B}}_k)>0}}{N_n(\tilde{\mathcal{B}}_k)}\Big|\tilde{\mathcal{B}}_k\right]+\frac{2^{2k+2}}{\pi^2}\mathbb{E}\left[\frac{\mathbb{1}_{N_n(\tilde{\mathcal{B}}_k)>0}}{N_n(\tilde{\mathcal{B}}_k)^2}\Big|\tilde{\mathcal{B}}_k\right]$$
$$+2^{k/2+1}\left(p-\frac{1}{2}\right)\mathbb{E}\left[\frac{\mathbb{1}_{N_n(\tilde{\mathcal{B}}_k)>0}}{N_n(\tilde{\mathcal{B}}_k)^{1/2}}\Big|\tilde{\mathcal{B}}_k\right]+\frac{2^{\frac{3k}{2}+2}}{\pi}\mathbb{E}\left[\frac{\mathbb{1}_{N_n(\tilde{\mathcal{B}}_k)>0}}{N_n(\tilde{\mathcal{B}}_k)^{3/2}}\Big|\tilde{\mathcal{B}}_k\right]$$

The computation is similar to (62), with $p'' = \mathbb{P}\left(X \in \tilde{\mathcal{B}}_k \mid \tilde{\mathcal{B}}_k\right)$ :

$$\begin{aligned} I_n + J_n &\leq \left(p - \frac{1}{2}\right)^2 + \frac{2^{k/2+3}(p-\frac{1}{2})}{\sqrt{\pi n \cdot p''}} + \left(2^k + 1 + \frac{2^{k+2}}{\pi}\left(p - \frac{1}{2}\right) + \frac{2^{3k/2+2}}{\pi} + \frac{2^{2k+2}}{\pi^2}\right)\frac{2}{(n+1)p''} \\ &\leq \left(p - \frac{1}{2}\right)^2 + \frac{2^{k/2+3}(p-\frac{1}{2})}{\sqrt{\pi n \cdot p''}} + \frac{2^{2k+3}}{\pi^2(n+1)p''}(1+\varepsilon_2(k)) \end{aligned}$$

with $\varepsilon_2(k) = \frac{\pi^2}{2^{(2k+3)}}\left(2^k + 1 + \frac{2^{k+2}}{\pi}(p-1/2) + \frac{2^{3k/2+2}}{\pi}\right)$. Finally,

$$\begin{aligned} &\mathbb{E}\left[(r_{k,1,n}(X) - (1-p))^2 \mathbb{1}_{X\in\mathcal{W}} \mathbb{1}_{X\in\tilde{\mathcal{B}}_k}\right] + \mathbb{E}\left[(r_{k,1,n}(X) - (1-p))^2 \mathbb{1}_{X\in\mathcal{W}} \mathbb{1}_{X\in\tilde{\mathcal{W}}_k}\right] \\ \leq &\mathbb{E}\left[(I'_n + J'_n)\mathbb{P}\left(X \in \mathcal{W}, X \in \tilde{\mathcal{B}}_k | \tilde{\mathcal{B}}_k\right)\right] + \left(p - \frac{1}{2}\right)^2 \mathbb{P}\left(X \in \mathcal{W}, X \in \tilde{\mathcal{W}}_k\right) \\ \leq &\mathbb{E}\left[\left(\left(p - \frac{1}{2}\right)^2 + \frac{2^{k/2+3}(p-\frac{1}{2})}{\sqrt{\pi n \cdot p''}} + \frac{2^{2k+3}}{\pi^2(n+1)p''}(1+\varepsilon_2(k))\right)\mathbb{P}\left(X \in \mathcal{W}, X \in \tilde{\mathcal{B}}_k | \tilde{\mathcal{B}}_k\right)\right] + \left(p - \frac{1}{2}\right)^2 \mathbb{P}\left(X \in \mathcal{W}, X \in \tilde{\mathcal{W}}\right. \end{aligned}$$

Since for all $\tilde{\mathcal{B}}_k$, there is exactly the same number of black cells and white cells in $\tilde{\mathcal{B}}_k$, we have

$$\mathbb{P}\left(X \in \mathcal{W}, X \in \tilde{\mathcal{B}}_k | \tilde{\mathcal{B}}_k\right) = \frac{p''}{2},$$

yielding

$$\begin{aligned} &\mathbb{E}\left[(r_{k,1,n}(X) - (1-p))^2 \mathbb{1}_{X\in\mathcal{W}} \mathbb{1}_{X\in\tilde{\mathcal{B}}_k}\right] + \mathbb{E}\left[(r_{k,1,n}(X) - (1-p))^2 \mathbb{1}_{X\in\mathcal{W}} \mathbb{1}_{X\in\tilde{\mathcal{W}}_k}\right] \qquad (76) \\ &\qquad \leq \frac{1}{2}\left(p - \frac{1}{2}\right)^2 + \frac{2^{k/2+2}(p-\frac{1}{2})}{\sqrt{\pi n}} + \frac{2^{2k+3}}{2\cdot\pi^2(n+1)}(1+\varepsilon_2(k)). \qquad (77) \end{aligned}$$

Gathering (42), (64) and (77), we have

$$\mathbb{E}\left[(r_{k,1,n}(X) - r(X))^2\right] \leq \left(p - \frac{1}{2}\right)^2 + \frac{2^{k/2+3}(p-\frac{1}{2})}{\sqrt{\pi n}} + \frac{7\cdot 2^{2k+2}}{\pi^2(n+1)}(1+\varepsilon(k)) + \frac{p^2+(1-p)^2}{2}\left(1 - \frac{1}{2^k}\right)^n$$

where $\varepsilon(k) = \frac{6\varepsilon_1(k)+\varepsilon_2(k)}{7}$.

## S5.2 Proof of 1. (ii) (lower-bound for the case $k < k^\star$)

We have, according to (42),

$$\begin{aligned} \mathbb{E}\left[(r_{k,1,n}(X) - r(X))^2\right] &= \mathbb{E}\left[(r_{k,1,n}(X) - r(X))^2 \mathbb{1}_{N_n(C_n(X)>0)}\right] + \mathbb{E}\left[(r(X))^2 \mathbb{1}_{N_n(C_n(X)=0)}\right] \\ &= \mathbb{E}\left[(r_{k,1,n}(X) - r(X))^2 \mathbb{1}_{N_n(C_n(X)>0)}\right] + \frac{p^2+(1-p)^2}{2}\mathbb{P}\left(N_n(C_n(X) = 0\right). \qquad (78) \end{aligned}$$

Letting $Z_2 = \mathbb{E}\left[\sum_{X_i \in C_n(X)} Y_i \mid N_1, ..., N_{2^k}, C_n(X)\right]$, we have

$$\mathbb{E}\left[(r_{k,1,n}(X) - r(X))^2 \mathbb{1}_{N_n(C_n(X)>0)}\right] \tag{79}$$

$$= \mathbb{E}\left[\left(\frac{\mathbb{1}_{N_n(C_n(X)>0)}}{N_n(C_n(X))} \sum_{X_i \in C_n(X)} Y_i - r(X)\right)^2 \mathbb{1}_{N_n(C_n(X)>0)}\right] \tag{80}$$

$$= \mathbb{E}\left[\frac{\mathbb{1}_{N_n(C_n(X)>0)}}{N_n(C_n(X))^2} \mathbb{E}\left[\left(\sum_{X_i \in C_n(X)} Y_i - N_n(C_n(X))r(X)\right)^2 \mid N_1, ..., N_{2^k}, C_n(X)\right]\right] \tag{81}$$

$$= \mathbb{E}\left[\frac{\mathbb{1}_{N_n(C_n(X)>0)}}{N_n(C_n(X))^2} \mathbb{E}\left[\left(\sum_{X_i \in C_n(X)} Y_i - Z_2\right)^2 + (Z_2 - N_n(C_n(X))r(X))^2 \right.\right. \tag{82}$$

$$\left.\left. + 2\left(\sum_{X_i \in C_n(X)} Y_i - Z_2\right)(Z_2 - N_n(C_n(X))r(X)) \mid N_1, ..., N_{2^k}, C_n(X)\right]\right]. \tag{83}$$

The cross-term is null according to the definition of $Z$ and because $(Z_2 - N_n(C_n(X)))$ is $(N_1, ..., N_{2^k}, C_n(X))$ - measurable. Therefore,

$$\mathbb{E}\left[\left(\frac{\mathbb{1}_{N_n(C_n(X)>0)}}{N_n(C_n(X))} \sum_{X_i \in C_n(X)} Y_i - r(X)\right)^2 \mathbb{1}_{N_n(C_n(X)>0)}\right] \tag{84}$$

$$= \mathbb{E}\left[\frac{\mathbb{1}_{N_n(C_n(X)>0)}}{N_n(C_n(X))^2} \mathbb{E}\left[\left(\sum_{X_i \in C_n(X)} Y_i - Z_2\right)^2 \mid N_1, ..., N_{2^k}, C_n(X)\right]\right] \tag{85}$$

$$+ \mathbb{E}\left[\frac{\mathbb{1}_{N_n(C_n(X)>0)}}{N_n(C_n(X))^2} \mathbb{E}\left[(Z_2 - N_n(C_n(X))r(X))^2 \mid N_1, ..., N_{2^k}, C_n(X)\right]\right]$$

$$= I_n + J_n, \tag{86}$$

where $I_n$ and $J_n$ are respectively a variance and bias term. Now, note that

$$\mathbb{E}\left[(Z_2 - N_n(C_n(X))r(X))^2 \mid N_1, ..., N_{2^k}, C_n(X)\right]$$

$$= \mathbb{E}\left[(Z_2 - N_n(C_n(X))p)^2 \mathbb{1}_{X \in \mathcal{B}} + (Z_2 - N_n(C_n(X))(1-p))^2 \mathbb{1}_{X \in \mathcal{W}} \mid N_1, ..., N_{2^k}, C_n(X)\right]. \tag{87}$$

Additionally,

$$\mathbb{P}(X \in \mathcal{B} \mid N_1, ..., N_{2^k}, C_n(X)) = \mathbb{P}(X \in \mathcal{W} \mid N_1, ..., N_{2^k}, C_n(X)) = 1/2.$$

Consequently,

$$\mathbb{E}\left[(Z_2 - N_n(C_n(X))r(X))^2 \mid N_1, ..., N_{2^k}, C_n(X)\right]$$

$$= \frac{1}{2}\mathbb{E}\left[(Z_2 - N_n(C_n(X))p)^2 + (Z_2 - N_n(C_n(X))(1-p))^2 \mid N_1, ..., N_{2^k}, C_n(X)\right]. \tag{88}$$

A small computation shows that for all $x \in \mathbb{R}$, for all $N \in \mathbb{N}$

$$(x - Np)^2 + (x - N(1-p))^2 \geq 2N^2(p - \frac{1}{2})^2,$$

which leads to

$$J_n \geq \left(p - \frac{1}{2}\right)^2 \mathbb{P}(N_n(C_n(X)) > 0).$$

All in all,

$$\mathbb{E}\left[(r_{k,1,n}(X)-r(X))^2\right] = I_n + J_n + \frac{p^2+(1-p)^2}{2}\mathbb{P}\left(N_n(C_n(X))=0\right) \tag{89}$$

$$\geq \left(p-\frac{1}{2}\right)^2 \mathbb{P}\left(N_n(C_n(X))>0\right) + \frac{p^2+(1-p)^2}{2}\mathbb{P}\left(N_n(C_n(X))=0\right) \tag{90}$$

$$\geq \left(p-\frac{1}{2}\right)^2. \tag{91}$$

## S5.3 Proof of 2. (i) (upper-bound for the case $k \geq k^\star$)

Recall that $k \geq k^\star$. In this case, each leaf of the encoding tree is included in a chessboard cell. Using (42), one gets

$$\mathbb{E}\left[(r_{k,1,n}(X)-r(X))^2)\right] = \mathbb{E}\left[(r_{k,1,n}(X)-r(X))^2 \mathbb{1}_{N_n(C_n(X))>0}\right] + \frac{p^2+(1-p)^2}{2}\left(1-\frac{1}{2^k}\right)^n. \tag{92}$$

Note that

$$\begin{aligned}
&\mathbb{E}\left[(r_{k,1,n}(X)-r(X))^2 \mathbb{1}_{N_n(C_n(X))>0}\right] \\
&= \mathbb{E}\left[\left(\frac{1}{N_n(\tilde{\mathcal{B}}_k)}\sum_{X_i\in\tilde{\mathcal{B}}_k} Y_i - p\right)^2 \mathbb{1}_{X\in\mathcal{B}}\mathbb{1}_{X\in\tilde{\mathcal{B}}_k}\right] + \mathbb{E}\left[\left(\frac{1}{N_n(\tilde{\mathcal{W}}_k)}\sum_{X_i\in\tilde{\mathcal{W}}_k} Y_i - p\right)^2 \mathbb{1}_{X\in\mathcal{B}}\mathbb{1}_{X\in\tilde{\mathcal{W}}_k}\right] \\
&+ \mathbb{E}\left[\left(\frac{1}{N_n(\tilde{\mathcal{B}}_k)}\sum_{X_i\in\tilde{\mathcal{B}}_k} Y_i - (1-p)\right)^2 \mathbb{1}_{X\in\mathcal{W}}\mathbb{1}_{X\in\tilde{\mathcal{B}}_k}\right] + \mathbb{E}\left[\left(\frac{1}{N_n(\tilde{\mathcal{W}}_k)}\sum_{X_i\in\tilde{\mathcal{W}}_k} Y_i - (1-p)\right)^2 \mathbb{1}_{X\in\mathcal{W}}\mathbb{1}_{X\in\tilde{\mathcal{W}}_k}\right] \\
&\leq \frac{1}{2}\mathbb{E}\left[\left(\frac{1}{N_n(\tilde{\mathcal{B}}_k)}\sum_{X_i\in\tilde{\mathcal{B}}_k} Y_i - p\right)^2 \mathbb{1}_{N_n(\tilde{\mathcal{B}}_k)>0}\right] + \frac{1}{2}\mathbb{E}\left[\left(\frac{1}{N_n(\tilde{\mathcal{W}}_k)}\sum_{X_i\in\tilde{\mathcal{W}}_k} Y_i - (1-p)\right)^2 \mathbb{1}_{N_n(\tilde{\mathcal{W}}_k)>0}\right] \\
&\qquad + \mathbb{E}\left[\mathbb{1}_{X\in\mathcal{B},X\in\tilde{\mathcal{W}}_k}\right] + \mathbb{E}\left[\mathbb{1}_{X\in\mathcal{W},X\in\tilde{\mathcal{B}}_k}\right].
\end{aligned} \tag{93}$$

Let L be a generic cell. The third term in (93) can be upper-bounded as follows:

$$\mathbb{E}\left[\mathbb{1}_{X\in\mathcal{B},X\in\tilde{\mathcal{W}}_k}\right] = \sum_{j=1}^{2^k}\mathbb{E}\left[\mathbb{1}_{X\in L_j}\mathbb{1}_{L_j\subset\tilde{\mathcal{W}}_k\cap\mathcal{B}}\right] \tag{94}$$

$$= \sum_{j=1}^{2^k}\mathbb{P}\left(X\in L_j\right)\mathbb{P}\left(L_j\subset\tilde{\mathcal{W}}_k\cap\mathcal{B}\right) \tag{95}$$

$$= \sum_{j=1}^{2^k}\mathbb{P}\left(X\in L_j\right)\mathbb{P}\left(L_j\subset\tilde{\mathcal{W}}_k \mid L_j\subset\mathcal{B}\right)\mathbb{P}\left(L_j\subset\mathcal{B}\right) \tag{96}$$

$$= \frac{1}{2}\mathbb{P}\left(L\subset\tilde{\mathcal{W}}_k \mid L\subset\mathcal{B}\right), \tag{97}$$

by symmetry. Now,

$$\mathbb{P}\left(L \subset \tilde{\mathcal{W}}_k \mid L \subset \mathcal{B}\right) = \mathbb{P}\left(\frac{1}{N_n(L)} \sum_{X_i \in L} \mathbb{1}_{Y_i=0} > \frac{1}{2} \mid L \subset \mathcal{B}\right) \tag{98}$$

$$\leq \mathbb{E}\left[\mathbb{P}\left(\frac{1}{N_n(L)} \sum_{X_i \in L, L \subset \mathcal{B}} \mathbb{1}_{Y_i=0} - (1-p) \geq \frac{1}{2} - (1-p) | N_n(L), L \subset \mathcal{B}\right) \mid L \subset \mathcal{B}\right] \tag{99}$$

$$\leq \mathbb{E}\left[e^{-2N_n(L)(p-\frac{1}{2})^2}\right] \tag{100}$$

(according to Hoeffding's inequality)

$$= \prod_{i=1}^{n} \mathbb{E}\left[e^{-2(p-\frac{1}{2})^2 \mathbb{1}_{X_i \in L}}\right] \tag{101}$$

(by independence of $X_i$'s)

$$= \left(\frac{e^{-2(p-\frac{1}{2})^2}}{2^k} + 1 - \frac{1}{2^k}\right)^n. \tag{102}$$

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

$$\mathbb{E}\left[\mathbb{1}_{X \in \mathcal{W}, X \in \tilde{\mathcal{B}}_k}\right] = \frac{1}{2}\mathbb{P}\left(L \subset \tilde{\mathcal{B}}_k \mid L \subset \mathcal{W}\right)$$
$$\leq \frac{1}{2}\left(\frac{e^{-2(p-\frac{1}{2})^2}}{2^k} + 1 - \frac{1}{2^k}\right)^n. \tag{103}$$

Therefore,

$$\mathbb{E}\left[(r_{k,1,n}(X) - r(X))^2)\right]$$
$$\leq \frac{1}{2}\mathbb{E}\left[\left(\frac{1}{N_n(\tilde{\mathcal{B}}_k)} \sum_{X_i \in \tilde{\mathcal{B}}_k} Y_i - p\right)^2 \mathbb{1}_{N_n(\tilde{\mathcal{B}}_k)>0}\right] + \frac{1}{2}\mathbb{E}\left[\left(\frac{1}{N_n(\tilde{\mathcal{W}}_k)} \sum_{X_i \in \tilde{\mathcal{W}}_k} Y_i - (1-p)\right)^2 \mathbb{1}_{N_n(\tilde{\mathcal{W}}_k)>0}\right]$$
$$+ \left(\frac{e^{-2(p-\frac{1}{2})^2}}{2^k} + 1 - \frac{1}{2^k}\right)^n + \frac{p^2 + (1-p)^2}{2}\left(1 - \frac{1}{2^k}\right)^n. \tag{104}$$

Now, the first term in (104) can be written as

$$\mathbb{E}\left[\left(\frac{1}{N_n(\tilde{\mathcal{B}}_k)} \sum_{X_i \in \tilde{\mathcal{B}}_k} Y_i - p\right)^2 \mathbb{1}_{N_n(\tilde{\mathcal{B}}_k)>0}\right] \tag{105}$$

$$= \mathbb{E}\left[\left(\frac{1}{N_n(\tilde{\mathcal{B}}_k)} \sum_{X_i \in \tilde{\mathcal{B}}_k} Y_i - p\right)^2 \mathbb{1}_{N_n(\tilde{\mathcal{B}}_k)>0} \mathbb{1}_{\mathcal{B}=\tilde{\mathcal{B}}_k}\right] + \mathbb{E}\left[\left(\frac{1}{N_n(\tilde{\mathcal{B}}_k)} \sum_{X_i \in \tilde{\mathcal{B}}_k} Y_i - p\right)^2 \mathbb{1}_{N_n(\tilde{\mathcal{B}}_k)>0} \mathbb{1}_{\mathcal{B} \neq \tilde{\mathcal{B}}_k}\right] \tag{106}$$

$$\leq \mathbb{E}\left[\left(\frac{1}{N_n(\tilde{\mathcal{B}}_k)} \sum_{X_i \in \tilde{\mathcal{B}}_k} Y_i - p\right)^2 \mathbb{1}_{N_n(\tilde{\mathcal{B}}_k)>0} \mathbb{1}_{\mathcal{B}=\tilde{\mathcal{B}}_k}\right] + \mathbb{P}\left(\mathcal{B} \neq \tilde{\mathcal{B}}_k\right) \tag{107}$$

Now, using a union bound, we obtain

$$\mathbb{P}\left(\mathcal{B} \neq \tilde{\mathcal{B}}_k\right) \leq \sum_{L_j \subset \mathcal{B}} \mathbb{P}\left(L_j \not\subset \tilde{\mathcal{B}}_k\right) + \sum_{L_j \subset \mathcal{W}} \mathbb{P}\left(L_j \subset \tilde{\mathcal{B}}_k\right) \tag{108}$$

$$\leq \frac{2^k}{2} \cdot \mathbb{P}\left(L \not\subset \tilde{\mathcal{B}}_k \mid L \subset \mathcal{B}\right) + \frac{2^k}{2} \cdot \mathbb{P}\left(L \subset \tilde{\mathcal{B}}_k \mid L \subset \mathcal{W}\right) \tag{109}$$

$$\leq 2^k \left(\frac{e^{-2(p-\frac{1}{2})^2}}{2^k} + 1 - \frac{1}{2^k}\right)^n, \tag{110}$$

according to (102) and (103). Additionally, the left term in (107) satisfies

$$\mathbb{E}\left[\left(\frac{1}{N_n(\tilde{\mathcal{B}}_k)} \sum_{X_i \in \tilde{\mathcal{B}}_k} Y_i - p\right)^2 \mathbb{1}_{N_n(\tilde{\mathcal{B}}_k)>0} \mathbb{1}_{\mathcal{B}=\tilde{\mathcal{B}}_k}\right] \leq \mathbb{E}\left[\left(\frac{1}{N_n(\mathcal{B})} \sum_{X_i \in \mathcal{B}} Y_i - p\right)^2 \mathbb{1}_{N_n(\mathcal{B})>0}\right] \tag{111}$$

$$\leq \mathbb{E}\left[\frac{\mathbb{1}_{N_n(\mathcal{B})>0}}{N_n(\mathcal{B})^2}\left(\sum_{X_i \in \mathcal{B}} Y_i - pN_n(\mathcal{B})\right)^2\right] \tag{112}$$

$$= p(1-p)\mathbb{E}\left[\frac{\mathbb{1}_{N_n(\mathcal{B})>0}}{N_n(\mathcal{B})}\right], \tag{113}$$

noticing that the square term of (112) is nothing but the conditional variance of a binomial distribution $B(N_n(\mathcal{B}), p)$. By Lemma S4 (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

$$\mathbb{E}\left[\left(\frac{1}{N_n(\tilde{\mathcal{B}}_k)} \sum_{X_i \in \tilde{\mathcal{B}}_k} Y_i - p\right)^2 \mathbb{1}_{\mathcal{B}=\tilde{\mathcal{B}}_k}\right] \leq \frac{2p(1-p)}{n+1} + 2^k \left(\frac{e^{-2(p-\frac{1}{2})^2}}{2^k} + 1 - \frac{1}{2^k}\right)^n. \tag{114}$$

Similarly,

$$\mathbb{E}\left[\left(\frac{1}{N_n(\tilde{\mathcal{W}}_k)} \sum_{X_i \in \tilde{\mathcal{W}}_k} Y_i - (1-p)\right)^2 \mathbb{1}_{N_n(\tilde{\mathcal{W}}_k)>0}\right] \leq \frac{2p(1-p)}{n+1} + 2^k \left(\frac{e^{-2(p-\frac{1}{2})^2}}{2^k} + 1 - \frac{1}{2^k}\right)^n. \tag{115}$$

Finally,

Injecting (114) and (115) into (104), 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.

## S5.4 Proof of 2. (ii) (lower-bound for the case $k \geq k^\star$)

We have

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

where

$$
\begin{aligned}
&\mathbb{E}\left[(r_{k,1,n}(X)-r(X))^2\mathbb{1}_{N_n(C_n(X))>0}\right] \\
&\geq \mathbb{E}\left[\left(\frac{1}{N_n(\tilde{\mathcal{B}}_k)}\sum_{X_i\in\tilde{\mathcal{B}}_k}Y_i-p\right)^2\mathbb{1}_{X\in\mathcal{B}}\mathbb{1}_{X\in\tilde{\mathcal{B}}_k}\mathbb{1}_{N_n(\tilde{\mathcal{B}}_k)>0}\mathbb{1}_{\mathcal{B}=\tilde{\mathcal{B}}_k}\right] \\
&\quad+\mathbb{E}\left[\left(\frac{1}{N_n(\tilde{\mathcal{W}}_k)}\sum_{X_i\in\tilde{\mathcal{W}}_k}Y_i-(1-p)\right)^2\mathbb{1}_{X\in\mathcal{W}}\mathbb{1}_{X\in\tilde{\mathcal{W}}_k}\mathbb{1}_{N_n(\tilde{\mathcal{W}}_k)>0}\mathbb{1}_{\mathcal{W}=\tilde{\mathcal{W}}_k}\right] \\
&\geq \mathbb{P}\left(X\in\mathcal{B}\right)\mathbb{E}\left[\left(\frac{1}{N_n(\tilde{\mathcal{B}}_k)}\sum_{X_i\in\tilde{\mathcal{B}}_k}Y_i-p\right)^2\mathbb{1}_{\mathcal{B}=\tilde{\mathcal{B}}_k}\mathbb{1}_{N_n(\tilde{\mathcal{B}}_k)>0}\right] \\
&\quad+\mathbb{P}\left(X\in\mathcal{W}\right)\mathbb{E}\left[\left(\frac{1}{N_n(\tilde{\mathcal{W}}_k)}\sum_{X_i\in\tilde{\mathcal{W}}_k}Y_i-(1-p)\right)^2\mathbb{1}_{\mathcal{W}=\tilde{\mathcal{W}}_k}\mathbb{1}_{N_n(\tilde{\mathcal{W}}_k)>0}\right]. \qquad (116)
\end{aligned}
$$

The first expectation term line (116) can be written as

$$
\mathbb{E}\left[\left(\frac{1}{N_n(\tilde{\mathcal{B}}_k)}\sum_{X_i\in\tilde{\mathcal{B}}_k}Y_i-p\right)^2\mathbb{1}_{\mathcal{B}=\tilde{\mathcal{B}}_k}\mathbb{1}_{N_n(\tilde{\mathcal{B}}_k)>0}\right]=\mathbb{P}\left(\mathcal{B}=\tilde{\mathcal{B}}_k\right)\mathbb{E}\left[\left(\frac{1}{N_n(\mathcal{B})}\sum_{X_i\in\mathcal{B}}Y_i-p\right)^2|\mathcal{B}=\tilde{\mathcal{B}}_k\right] \qquad (117)
$$

According to (110),

$$
\mathbb{P}\left(\mathcal{B}=\tilde{\mathcal{B}}_k\right)\geq 1-2^k\cdot\left(1+\frac{e^{-2(p-\frac{1}{2})^2}-1}{2^k}\right)^n. \qquad (118)
$$

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,

$$
\mathbb{E}\left[\left(\frac{1}{N_n(\mathcal{B})}\sum_{X_i\in\mathcal{B}}Y_i-p\right)^2|\mathcal{B}=\tilde{\mathcal{B}}_k\right]=\mathbb{E}\left[\frac{1}{N_n(\mathcal{B})^2}\mathbb{E}\left[\left(\sum_{X_i\in\mathcal{B}}Y_i-N_n(\mathcal{B})p\right)^2|N_1,...N_{2^k},\mathcal{B}=\tilde{\mathcal{B}}_k\right]|\mathcal{B}=\tilde{\mathcal{B}}_k\right] \qquad (119)
$$

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

$$\mathbb{E}\left[\left(\sum_{X_i \in \mathcal{B}} Y_i - N_n(\mathcal{B})p\right)^2 | N_1, ...N_{2^k}, \mathcal{B} = \tilde{\mathcal{B}}_k\right] \tag{120}$$

$$= \mathbb{E}\left[\left(Z - \mathbb{E}\left[Z \mid N_1, ...N_{2^k}, \tilde{\mathcal{B}}_k = \mathcal{B}\right]\right)^2 + \left(\mathbb{E}\left[Z \mid N_1, ...N_{2^k}, \tilde{\mathcal{B}}_k = \mathcal{B}\right] - N_n(\mathcal{B})p\right)^2 \mid N_1, ...N_{2^k}, \tilde{\mathcal{B}}_k = \mathcal{B}\right] \tag{121}$$

$$\geq \mathbb{E}\left[\left(Z - \mathbb{E}\left[Z \mid N_1, ...N_{2^k}, \tilde{\mathcal{B}}_k = \mathcal{B}\right]\right)^2 \mid N_1, ...N_{2^k}, \tilde{\mathcal{B}}_k = \mathcal{B}\right] \tag{122}$$

$$= \mathbb{E}\left[\left(\sum_{L_j \subset \mathcal{B}} Z_j - \mathbb{E}\left[Z_j \mid N_j, L_j \subset \tilde{\mathcal{B}}_k\right]\right)^2 \mid N_1, ...N_{2^k}, \tilde{\mathcal{B}}_k = \mathcal{B}\right] \tag{123}$$

$$= \sum_{L_j \subset \mathcal{B}} \mathbb{E}\left[\left(Z_j - \mathbb{E}\left[Z_j \mid N_j, L_j \subset \tilde{\mathcal{B}}_k\right]\right)^2 \mid N_j, L_j \subset \tilde{\mathcal{B}}_k\right]$$
$$+ 2 \sum_{L_i, L_j \subset \mathcal{B}, L_i \neq L_j} \mathbb{E}\left[\left(Z_i - \mathbb{E}\left[Z_i \mid N_i, L_i \subset \tilde{\mathcal{B}}_k\right]\right)\left(Z_j - \mathbb{E}\left[Z_j \mid N_j, L_j \subset \tilde{\mathcal{B}}_k\right]\right) \mid N_1, ...N_{2^k}, \tilde{\mathcal{B}}_k = \mathcal{B}\right] \tag{124}$$

$$= \sum_{L_j \subset \mathcal{B}} \mathbb{E}\left[\left(Z_j - \mathbb{E}\left[Z_j \mid N_j, L_j \subset \tilde{\mathcal{B}}_k\right]\right)^2 \mid N_j, L_j \subset \tilde{\mathcal{B}}_k\right]. \tag{125}$$

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 (123) and (124), 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 (125), 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 $\{1, ..., 2^k\}$,

$$\mathbb{E}\left[\left(Z_j - \mathbb{E}\left[Z_j \mid N_j, L_j \subset \tilde{\mathcal{B}}_k\right]\right)^2 \mid N_j, L_j \subset \tilde{\mathcal{B}}_k\right] \tag{126}$$

$$= \mathbb{E}\left[Z_j^2 \mid N_j, L_j \subset \tilde{\mathcal{B}}_k\right] - \mathbb{E}\left[Z_j \mid N_j, L_j \subset \tilde{\mathcal{B}}_k\right]^2 \tag{127}$$

$$\geq \mathbb{E}\left[Z_j^2 \mid N_j\right] - \mathbb{E}\left[Z_j \mid N_j, L_j \subset \tilde{\mathcal{B}}_k\right]^2 \tag{128}$$

$$= N_j p(1-p) + N_j^2 p^2 - \left(N_j p + \frac{N_j}{2}(1-p)\frac{\mathbb{P}\left(Z_j = \frac{N_j}{2} \mid N_j\right)}{\sum_{i=\frac{N_j}{2}}^{N_j} \mathbb{P}\left(Z_j = i\right)}\right)^2 \tag{129}$$

$$\geq N_j(1-p)\left(p - N_j(1-p)\mathbb{P}\left(Z_j = \frac{N_j}{2} \mid N_j\right)^2 - 2N_j p \cdot \mathbb{P}\left(Z_j = \frac{N_j}{2} \mid N_j\right)\right) \tag{130}$$

$$\geq N_j(1-p)\left(p - \frac{N_j(1-p)}{\pi\left(\frac{N_j}{2} + \frac{1}{4}\right)}\left(4p(1-p)\right)^{N_j} - \frac{2N_j}{\sqrt{\pi\left(\frac{N_j}{2} + \frac{1}{4}\right)}}\left(4p(1-p)\right)^{N_j/2}\right) \tag{131}$$

$$\geq N_j p(1-p) - \left(\frac{2(1-p)^2}{\pi} + 2\sqrt{2}(1-p)\right) \cdot N_j^{3/2} \cdot \left(4p(1-p)\right)^{N_j/2}. \tag{132}$$

We deduced Line (128) from the fact that $Z_j^2$ is a positive random variable, (129) from Lemma (S4) (v), Line (130) from the fact that $p > 1/2$ and Line (131) from the inequality (3) on the binomial

coefficient. Injecting (124) and (132) into (119) yields

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

$$\geq\mathbb{E}\left[\frac{1}{N_n(\mathcal{B}_k)^2}\sum_{L_j\subset\mathcal{B}}\left(N_jp(1-p)-\left(\frac{2(1-p)^2}{\pi}+2\sqrt{2}(1-p)\right)\cdot N_j^{3/2}\cdot(4p(1-p))^{N_j/2}\right)|\mathcal{B}=\tilde{\mathcal{B}}_k\right] \tag{133}$$

$$\geq\mathbb{E}\left[\frac{p(1-p)}{N_n(\mathcal{B})}\mid\mathcal{B}=\tilde{\mathcal{B}}_k\right]-\left(\frac{2(1-p)^2}{\pi}+2\right)\sum_{L_j\subset\mathcal{B}}\mathbb{E}\left[(4p(1-p))^{N_j/2}\mid\mathcal{B}=\tilde{\mathcal{B}}_k\right] \tag{134}$$

$$\geq p(1-p)\mathbb{E}\left[\frac{1}{N_n(\mathcal{B})}\mid\mathcal{B}=\tilde{\mathcal{B}}_k\right]-3\cdot2^{k-1}\mathbb{E}\left[(4p(1-p))^{N_b/2}\mid\mathcal{B}=\tilde{\mathcal{B}}_k\right] \tag{135}$$

where the last inequality relies on the fact that the $N_j, L_j\subset\mathcal{B}$ are i.i.d, with $b\in1,...,2^k$ be the index of a cell included in $\mathcal{B}$. $N_j$ is a binomial random variable $\mathfrak{B}(n,2^{-k})$.

$$\mathbb{E}\left[(4p(1-p))^{N_j/2}\mid\mathcal{B}=\tilde{\mathcal{B}}_k\right]\leq\mathbb{E}\left[(4p(1-p))^{N_j/2}\right]\frac{1}{\mathbb{P}\left(\mathcal{B}=\tilde{\mathcal{B}}_k\right)} \tag{136}$$

$$=\left(\sqrt{4p(1-p)}\cdot2^{-k}+(1-2^{-k})\right)^n\frac{1}{\mathbb{P}\left(\mathcal{B}=\tilde{\mathcal{B}}_k\right)}. \tag{137}$$

From the inequality Line (118), we deduce that as soon as $n\geq\frac{(k+1)\log(2)}{\log(2^k)-\log(e^{-2(p-1/2)^2}-1+2^k)}$,

$$\frac{1}{\mathbb{P}\left(\mathcal{B}=\tilde{\mathcal{B}}_k\right)}\leq2. \tag{138}$$

Therefore,

$$\mathbb{E}\left[(4p(1-p))^{N_j/2}\mid\mathcal{B}=\tilde{\mathcal{B}}_k\right]\leq2\left(\sqrt{4p(1-p)}\cdot2^{-k}+(1-2^{-k})\right)^n. \tag{139}$$

Moreover,

$$\mathbb{E}\left[\frac{1}{N_n(\mathcal{B})}|\mathcal{B}=\tilde{\mathcal{B}}_k\right]\geq\frac{1}{\mathbb{E}\left[N_n(\mathcal{B})|\mathcal{B}=\tilde{\mathcal{B}}_k\right]} \tag{140}$$

$$\geq\frac{\mathbb{P}\left(\mathcal{B}=\tilde{\mathcal{B}}_k\right)}{\mathbb{E}\left[N_n(\mathcal{B})\right]} \tag{141}$$

$$\geq\frac{2}{n}-\frac{2^{k+1}}{n}\left(1+\frac{e^{-2(p-\frac{1}{2})^2}-1}{2^k}\right)^n \tag{142}$$

where the last inequality comes from the probability bound line (118) and the fact that $N_n(\mathcal{B})$ is a binomial random variable $\mathfrak{B}(n,1/2)$.

Finally,

$$\mathbb{E}\left[\left(\frac{1}{N_n(\mathcal{B})}\sum_{X_i\in\mathcal{B}}Y_i-p\right)^2|\mathcal{B}=\tilde{\mathcal{B}}_k\right] \tag{143}$$

$$\geq\frac{2p(1-p)}{n}-3\cdot2^k\left(1-2^{-k}\left(1-\sqrt{4p(1-p)}\right)\right)^n-\frac{2^{k+1}p(1-p)}{n}\left(1+\frac{e^{-2(p-\frac{1}{2})^2}-1}{2^k}\right)^n. \tag{144}$$

Similarly, regarding the second term of (116), 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 :

$$\mathbb{E}\left[\left(\frac{1}{N_n(\mathcal{W})}\sum_{X_i\in\mathcal{W}} Y_i - p\right)^2 |\mathcal{W}=\tilde{\mathcal{W}}_k\right] \tag{145}$$

$$\geq \frac{2p(1-p)}{n} - 3\cdot 2^k\left(1-2^{-k}\left(1-\sqrt{4p(1-p)}\right)\right)^n - \frac{2^{k+1}p(1-p)}{n}\left(1+\frac{e^{-2(p-\frac{1}{2})^2}-1}{2^k}\right)^n. \tag{146}$$

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 S6.** *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,

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

□