Analyzing Bagging

Peter Bühlmann
ETH Zürich, Switzerland

Bin Yu
University of California at Berkeley

Revised Version
October 2001

Abstract

Bagging is one of the most effective computationally intensive procedures to improve on unstable estimators or classifiers, useful especially for high dimensional data set problems. Here we formalize the notion of instability and derive theoretical results to analyze the variance reduction effect of bagging (or its variant) in mainly hard decision problems, which include estimation after testing in regression and decision trees for continuous regression functions and classifiers. Hard decisions create instability, and bagging is shown to smooth such hard decisions yielding smaller variance and mean squared error. With theoretical explanations, we motivate subagging based on subsampling as an alternative aggregation scheme. It is computationally cheaper but still showing approximately the same accuracy as bagging. Moreover, our theory reveals improvements in first order and in line with simulation studies.

In particular, we obtain an asymptotic limiting distribution at the cube-root rate for the split point when fitting piecewise constant functions. Denoting sample size by n, it follows that in a cylindric neighborhood of diameter $n^{-1/3}$ of the theoretically optimal split point, the variance and mean squared error reduction of subagging can be characterized analytically. Because of the slow rate, our reasoning also provides an explanation on the global scale for the whole covariate space in a decision tree with finitely many splits.

Heading: Analyzing bagging

1 Introduction

Advances in data collection and computing technologies have led to the proliferation of large data sets. Bagging is one of the recent and successful computationally intensive methods for improving unstable estimation or classification schemes. It is extremely useful for large, high dimensional data set problems where finding a good model or classifier in one step is impossible because of the complexity and scale of the problem. Bagging [**b**ootstrap **agg**regat**ing**] was introduced by Breiman (1996a) to reduce the variance of a predictor. It has attracted much attention and is frequently applied, although deep theoretical insight has been lacking. Here we take a substantial step towards a better understanding of bagging and its variant subagging [**sub**sample **agg**regat**ing**].

Consider the regression set-up. The data is denoted by $L_i = (Y_i, X_i)$ $(i = 1, \ldots, n)$ with Y_i the real-valued response and X_i a p-dimensional explanatory variable for the i-th instance. Given a new explanatory feature or covariate x, a predictor for $\mathbb{E}[Y|X = x] = f(x)$ [or of the response variable corresponding to x] is denoted by

$$\hat{\theta}_n(x) = h_n(L_1, ..., L_n)(x).$$

This estimator could involve a complex model or learning algorithm, for example linear regression with variable selection via testing, regression trees such as CART (Breiman et al., 1984) or MARS (Friedman, 1991).

Definition 1.1 *[Bagging]. Theoretically, bagging is defined as follows.*

(I) Construct a bootstrap sample $L_i^* = (Y_i^*, X_i^*)$ $(i = 1, \ldots, n)$ according to the empirical distribution of the pairs $L_i = (Y_i, X_i)$ $(i = 1, \ldots, n)$.

(II) Compute the bootstrapped predictor $\hat{\theta}_n^*(x)$ by the plug-in principle; i.e., $\hat{\theta}_n^*(x) = h_n(L_1^*, \ldots, L_n^*)(x)$, where $\hat{\theta}_n(x) = h_n(L_1, \ldots, L_n)(x)$.

(III) The bagged predictor is $\hat{\theta}_{n;B}(x) = \mathbb{E}^*[\hat{\theta}_n^*(x)]$.

In practice, the bootstrap expectation in (III) is implemented by Monte Carlo: for every bootstrap simulation $j \in \{1, \ldots, J\}$ from (I), we compute $\hat{\theta}_{n;(j)}^*(x)$ $(j = 1, \ldots, J)$ as in (II) to approximate $\hat{\theta}_{n;B}(x) \approx J^{-1} \sum_{j=1}^{J} \hat{\theta}_{n;(j)}^*(x)$. J is often chosen in the range of 50, depending on sample size and on the computational cost to evaluate the predictor, see Breiman (1996a, section 6.2).

Breiman (1996a) describes heuristically the performance of bagging as follows. The variance of the bagged estimator $\hat{\theta}_{n;B}(x)$ is equal or smaller than that for the original estimator $\hat{\theta}_n(x)$. There can be a drastic variance reduction if the original predictor is 'unstable'. On the other hand, the magnitudes of the bias are roughly the same for the bagged and the original procedure. It implies that bagging improves the mean squared error a lot for 'unstable' predictors whereas it remains roughly the same for 'stable' schemes. This has been observed in empirical studies, cf. Breiman (1996a). We add here deeper insight based on theoretical results and correct some previous beliefs about bagging.

Breiman (1996b) gives a heuristic definition of instability: a predictor is 'unstable if small changes in the data can cause large changes in the predicted value(s). We formalize here a precise definition that is not inconsistent with Breiman's.

Definition 1.2 *[Stability of a predictor]. A statistic $\hat{\theta}_n(x) = h_n(L_1, \ldots, L_n)(x)$ is called stable if $\hat{\theta}_n = \theta + o_P(1)$ $(n \to \infty)$ for some fixed value θ.*

Although this definition resembles very much the one for consistency, it is very different since the value θ here is only a stable limit and not necessarily the parameter of interest. Instability thus takes place whenever the procedure $\hat{\theta}_n$ is not converging to a fixed value: another [even infinitely long] realization from the data generating distribution would produce a different value of the procedure, with positive probability. Much of our coverage of bagging will be on unstable predictors as defined above. They arise mainly when hard decisions with indicators are involved as in decision trees [see sections 2 and 3].

Theoretical investigations on why bagging works have been given by Friedman and Hall (2000) and Buja and Stuetzle (2000a, 2000b). Friedman and Hall (2000) decompose a smooth estimator into linear and higher orders. They argue heuristically that bagging reduces variance for the higher order terms, and the linear term remains unaffected. Buja and Stuetzle (2000a) concentrate on U-statistics to give a clear and rigorous answer about bagging: the leading effects of bagging on variance, squared bias and mean squared error are of second order n^{-2}. Thus, bagging potentially improves mean squared error in the second [and higher] order asymptotic term only, but doesn't affect the leading first order term. Moreover, Buja and Stuetzle show that bagging sometimes even increases the second order MSE terms inducing a damaging effect. Despite of the fact that these previous works are nonlinear, they do not cover the prominent case of decision trees.

For non-smooth and unstable predictors, we demonstrate in this paper that bagging does improve the *first order* dominant variance and mean squared error asymptotic terms, as much as a factor of 3. Such prediction schemes include decision trees like CART and subset model selection techniques via testing, where indicators play a prominent role. We pay special attention to decision trees with one or finitely many binary splits, a so-called stump or best-first induced binary tree [without pruning], respectively. The asymptotics are non-standard: the splitting variable turns out to have a convergence rate $n^{-1/3}$ and the limiting distribution can only be characterized in terms of Airy functions [see Groeneboom, 1989] not leading to a closed [or at least 'simpler'] expression. In such non-standard problems, the bootstrap in the bagging procedure described above does not work in the conventional sense and is hard to analyze, at least from a theoretical point of view. As a promising variant of bagging, more accessible for analysis, we study *subagging* [**sub**sample **agg**regat**ing**] in section 3.2. But *unlike* more standard approaches to subsampling without replacement, we choose the subsample size $m = [an]$ with $0 < a < 1$. This has also appeared in Friedman and Hall (2000) and Buja and Stuetzle (2000a, 2000b). Based on rigorous results for subagged stumps and best-first induced decision trees with finitely many splits, we show that subagging improves upon variance and mean squared error. Besides theoretical arguments, subagging also has substantial *computational* advantages since the original predictor is only evaluated many times for m instead of n data points. Our results also illuminate why bagging combined with boosting [cf. Bühlmann and Yu, 2000] can be a very effective method achieving both variance *and* bias reduction for decision trees.

Unlike previously suggested, the success of bagging is not exclusively restricted to high-dimensional schemes, since it works also well for stumps which involve only three parameters [when the coordinate axis to split is assumed fixed]. Only Buja and Stuetzle

(2000b) also make this point that the properties of bagging are not primarily depending on dimensionality. When the original predictor involves a hard-thresholding indicator decision, our results show that bagging [and variants thereof] can be interpreted as some data-driven *soft-thresholding* schemes, which are characterized analytically. In order to compare with hard decision tree schemes like CART, we give a rigorous asymptotic result for the basic element in MARS [Friedman, 1991], as a prime example for a predictor involving a continuous, but non-smooth decision. There, bagging, or variants thereof, do not increase [substantially] the prediction performance.

The rest of the paper is organized as follows. Section 2 contains results for predictors, discontinuous and continuous, involving the conventional $n^{-1/2}$-convergence rate. Section 3 introduces subagging, gives the non-standard $n^{-1/3}$-rate result for the split in a binary tree, and explains the variance reduction effect of subagging for such trees. The theoretical arguments and interpretations are supported by some numerical experiments in section 4. Conclusions are given in section 5 and the more involved proofs are collected in section 6.

2 Bagging with indicators: the standard case

A linear predictor remains the same under bagging. The simplest example is

$$\hat{\theta}_n(x) \equiv \overline{Y}_n = n^{-1} \sum_{i=1}^{n} Y_i$$

with no explanatory variable x. Then

$$\overline{Y}_{n;B}(x) = \mathbb{E}^*[Y_1^*] = \overline{Y}_n.$$

Thus, the only interesting case has predictors $\hat{\theta}_n(x)$ that are nonlinear functions of the data. For $\hat{\theta}_n(x)$ being a U-statistic, Buja and Stuetzle (2000a) show that under some 'ideal' circumstances, bagging reduces the variance of the the higher order but *not* of the leading first order asymptotic term; they also show that bagging U-statistics may increase mean squared error, depending on the data-generating probability distribution.

A very different type of estimators is studied here: we consider non-differentiable, and even discontinuous predictors $\hat{\theta}_n(x)$ which cannot be easily expanded. The classical smooth function theory used by Friedman and Hall (2000) and Buja and Stuetzle (2000a, 2000b) does not apply. We particularly consider predictors involving indicator functions. They arise whenever a hard decision is made. Examples include CART as a decision tree or variable selection in regression models, for which most of the empirical success of bagging has been reported, cf. Breiman (1996a), Bauer and Kohavi (1999).

2.1 Plug-in applied to an indicator

One of the main ideas behind why bagging works can be demonstrated with a simple toy example. Consider the predictor

$$\hat{\theta}_n(x) = \mathbb{1}_{[\overline{Y}_n \le x]},\ x \in \mathbb{R},$$

where $\overline{Y}_n$, the average of the Y_i's based on data $L_i = (Y_i, X_i)$ $(i = 1, \ldots, n)$. For a fixed x, $\hat{\theta}_n(x)$ is the result of thresholding $\overline{Y}_n$ at the threshold x; for a fixed $\overline{Y}_n$, it is as a thresholding function of x at the threshold $\overline{Y}_n$.

Heuristically, bagging averages here over indicators [as functions of x] with thresholds varying around $\overline{Y}_n$ [since the bootstrapped $\overline{Y}_n^*$ fluctuates around $\overline{Y}_n$], resulting in a smoothed or soft indicator.

If we take the view of fixed x, after a proper scaling of x, a simple yet precise analysis below shows that bagging is a smoothing operation for thresholding $\overline{Y}_n$ at x. Due to the central limit theorem we have

$$n^{1/2}(\overline{Y}_n - \mu) \to_D \mathcal{N}(0, \sigma^2) \tag{2.1}$$

with $\mu = \mathbb{E}[Y_1]$ and $\sigma^2 = \mathrm{Var}(Y_1)$. Then for an x in the $n^{-1/2}$-neighborhood of the parameter μ

$$x = x_n(c) = \mu + c\sigma n^{-1/2}, \tag{2.2}$$

we have the distributional approximation

$$\hat{\theta}_n(x_n(c)) \overset{\mathcal{D}}{\approx} \mathbb{1}_{[Z \le c]},\ Z \sim \mathcal{N}(0, 1). \tag{2.3}$$

Obviously, for a fixed c, this is a hard decision function of Z, the limiting random quantity from the asymptotic distribution of $\overline{Y}_n$. Denoting by $\Phi(\cdot)$ the c.d.f. of a standard normal distribution, it follows that

$$\begin{aligned} &\mathbb{E}[\hat{\theta}_n(x_n(c))] \to \mathbb{P}[Z \le c] = \Phi(c)\ (n \to \infty), \\ &\mathrm{Var}(\hat{\theta}_n(x_n(c))) \to \Phi(c)(1 - \Phi(c))\ (n \to \infty). \end{aligned} \tag{2.4}$$

Since the variance does not converge to zero, $\hat{\theta}_n(x_n(c))$ is unstable in the sense of Definition 1.2: the predictor assumes the values 0 and 1 with a positive probability, even as n tends to infinity. On the other hand, averaging for the bagged predictor looks as follows:

$$\begin{aligned} \hat{\theta}_{n;B}(x_n(c)) &= \mathbb{E}^*[\mathbb{1}_{[\overline{Y}_n^* \le x_n(c)]}] = \mathbb{E}^*[\mathbb{1}_{[n^{1/2}(\overline{Y}_n^* - \overline{Y}_n)/\sigma \le n^{1/2}(x_n(c) - \overline{Y}_n)/\sigma]}] \\ &= \Phi(n^{1/2}(x_n(c) - \overline{Y}_n)) + o_P(1) \\ &\overset{\mathcal{D}}{\approx} \Phi(c - Z),\ Z \sim \mathcal{N}(0, 1), \end{aligned} \tag{2.5}$$

where the first approximation [second line] follows because the bootstrap works for the arithmetic mean $\overline{Y}_n$ [see (A1) below] and the second because of (2.1) and the definition of $x_n(c)$ in (2.2). Comparing with (2.3), bagging produces a soft decision function of Z: it is a shifted inverse probit, similar to a sigmoid-type function; see also Figure 2.1.

Bagging reduces variance due to the smoothing or soft- instead of hard-thresholding operation. An instructive case is with $x = x_n(0) = \mu$; i.e., x is exactly at the most unstable location, where $\mathrm{Var}(\hat{\theta}_n(x))$ is maximal. Formula (2.5) gives

$$\hat{\theta}_{n;B}(x_n(0)) \to_D \Phi(-Z) = U,\ U \sim \mathrm{Uniform}([0, 1]).$$

Thus,

$$\begin{aligned} &\mathbb{E}[\hat{\theta}_{n;B}(x_n(0))] \to \mathbb{E}[U] = 1/2\ (n \to \infty) \\ &\mathrm{Var}(\hat{\theta}_{n;B}(x_n(0))) \to \mathrm{Var}(U) = 1/12\ (n \to \infty). \end{aligned}$$

Comparing with (2.4), bagging is asymptotically unbiased [the asymptotic parameter to be estimated is $\lim_{n\to\infty}\mathbb{E}[\hat{\theta}_n(x_n(0))] = \Phi(0) = 1/2$], but the asymptotic variance is reduced by a factor 3! We will see later that for a whole range where $c \neq 0$ in (2.2) [i.e., $x \neq \mu$], bagging still reduces variance while adding only little to the bias.

We now look at a more general indicator predictor of similar form

$$\hat{\theta}_n(x) = \mathbb{1}_{[\hat{d}_n \leq x]},\ x \in \mathbb{R}, \tag{2.6}$$

where the threshold $\hat{d}_n$ satisfies the following assumption:

(A1) For some increasing sequence $(b_n)_{n\in\mathbb{N}}$ and the bootstrapped estimator $\hat{d}_n^*$, we have

$$b_n(\hat{d}_n - d^0) \to_D \mathcal{N}(0, \sigma_\infty^2),$$
$$\sup_{v\in\mathbb{R}} |\mathbb{P}^*[b_n(\hat{d}_n^* - \hat{d}_n) \leq v] - \Phi(v/\sigma_\infty)| = o_P(1),$$

with $0 < \sigma_\infty^2 < \infty$.

Our example with $\hat{d}_n = \overline{Y}_n$ satisfies this assumption with $b_n = n^{1/2}$. (A1) generally requires asymptotic normality of the estimator with any rate and that the bootstrap works. Due to the results in Giné and Zinn (1990), this essentially holds by assuming i.i.d. observations and $\hat{d}_n$ being a smooth functional evaluated at the empirical distribution.

Proposition 2.1 *Assume (A1). For the predictor in (2.6) with* $x = x_n(c) = d^0 + c\sigma_\infty b_n^{-1}$,

$$\hat{\theta}_n(x_n(c)) \to_D g(Z) = \mathbb{1}_{[Z\leq c]},$$
$$\hat{\theta}_{n;B}(x_n(c)) \to_D g_B(Z) = \Phi(c - Z),$$

where $Z \sim \mathcal{N}(0,1)$.

Proof: The distributional limits follow in exact analogy to (2.3) and (2.5). □

Figure 2.1 illustrates the two functions $g(\cdot)$ and $g_B(\cdot)$ from Proposition 2.1. Bagging

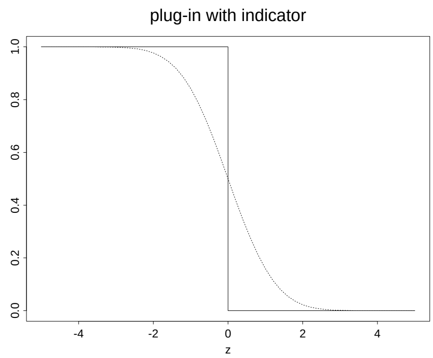

Figure 2.1: Indicator predictor from (2.6) at $x = x_n(0)$ as in (2.2) or Proposition 2.1. Function $g(z) = \mathbb{1}_{[z\leq 0]}$ [solid line] and $g_B(z)$ [dotted line] defining the asymptotics of the predictor and its bagged version [see Proposition 2.1].

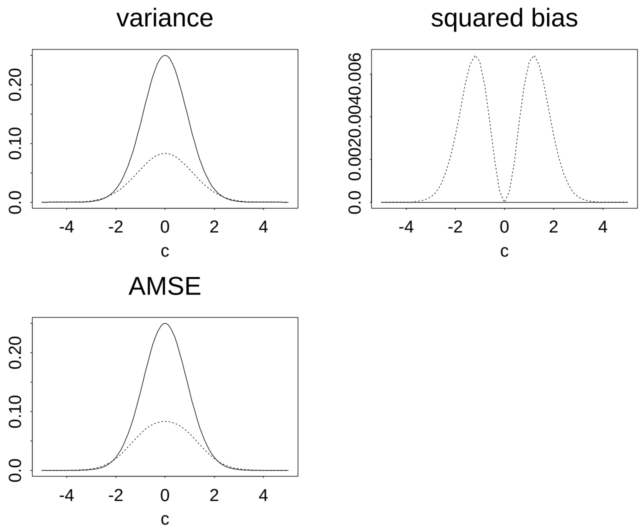

Figure 2.2: Indicator predictor from (2.6) at $x = x_n(c)$ as in (2.2) or Proposition 2.1. Asymptotic variance, squared bias and mean squared error [AMSE] for the predictor $\hat{\theta}_n(x_n(c))$ from (2.6) [solid line] and for the bagged predictor $\hat{\theta}_{n;B}(x_n(c))$ [dotted line] as a function of c.

reduces variance due to the smoothing or soft- instead of hard-thresholding operation. We compute the first two asymptotic moments in the unstable region with $x = x_n(c)$. Denote the convolution of f and g by $f * g(\cdot) = \int_{\mathbb{R}} f(\cdot - y)g(y)dy$, and the standard normal density by $\varphi(\cdot)$.

Corollary 2.1 *Assume (A1). For the predictor in (2.6) with $x = x_n(c)$ as in Proposition 2.1,*

(i) $\lim_{n\to\infty}\mathbb{E}[\hat{\theta}_n(x_n(c))] = \Phi(c)$,
$\lim_{n\to\infty} \mathrm{Var}(\hat{\theta}_n(x_n(c))) = \Phi(c)(1 - \Phi(c))$.

(ii) $\lim_{n\to\infty}\mathbb{E}[\hat{\theta}_{n;B}(x_n(c))] = \Phi * \varphi(c)$,
$\lim_{n\to\infty} \mathrm{Var}(\hat{\theta}_{n;B}(x_n(c))) = \Phi^2 * \varphi(c) - (\Phi * \varphi(c))^2$.

Proof: Assertion (i) is straightforward. Assertion (ii) follows by Proposition 2.1 together with the boundedness of the function $g_B(\cdot)$ therein. □

Numerical evaluations of these first two asymptotic moments and mean squared error [MSE] are given in Figure 2.2. We see that for $|c| \leq 2.3$, bagging improves the mean squared error. The biggest gain is at the most unstable point $x = d^0$, corresponding to $c = 0$. The squared bias with bagging has only a negligible effect on the MSE [note the different scales in Figure 2.2].

2.2 Variable selection via testing in linear models

In this section, we show that since estimation after testing in linear regression is a hard-thresholding operation, bagging also acts as smoothing or softening and leads to a reduced variance without much sacrifice on the bias.

Consider the linear model

$$Y_i = (\mathbf{X}\beta)_i + \varepsilon_i \ (i = 1, \ldots, n),$$

where $\mathbf{X}$ is the $n \times p$ random design matrix (X_{ij}), β is a $p \times 1$ parameter vector and $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. with expectation zero and variance σ^2. Assume that the columns in $\mathbf{X}$ are orthogonal [in expectation]: this simplifies the mathematical problem, although the results are expected to be relevant by weakening this requirement. The least squares estimate $\hat{\beta}_n$ is then asymptotically normally distributed at rate $n^{-1/2}$ [assuming finiteness of the second moment of the covariate vector] with independent components; testing individual hypotheses $H_{0,j}: \ \beta_j = 0 \ (j = 1, \ldots, p)$ is thus a reasonable model selection procedure. A predictor of interest is then

$$\hat{\theta}_n(x) = \sum_{j=1}^{p} \hat{\beta}_j \mathbb{1}_{[|\hat{\beta}_j| > u_{n,j}]} x^{(j)}$$

with $x^{(j)}$ the j-th component of x. For example, the thresholds could be $u_{n,j} = C_j n^{-1/2}$: the choice $C_j = t_{1-\alpha/2;n-1}\hat{\sigma}/\sqrt{n^{-1}\sum_{i=1}^{n} X_{ij}^2}$ would correspond to the [conditional] t-test on significance level α. Due to the asymptotic independence of the components of $\hat{\beta}$, the MSE is asymptotically additive with p individual MSEs. We thus consider without loss of

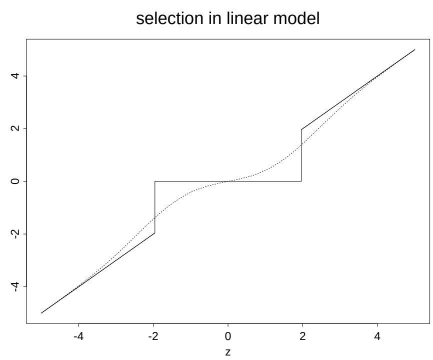

Figure 2.3: Top: predictor from (2.7), with $x = 1$ and u_n as in (2.9) with $c = 1.96$, in linear model (2.8). Function $g(z)$ [solid line] and $g_B(z)$ [dotted line] from Proposition 2.2, defining the asymptotics of $\hat{\theta}_n(1)$ and its bagged version, respectively.

generality the predictor

$$\hat{\theta}_n(x) = \hat{\beta}\mathbb{1}_{[|\hat{\beta}| > u_n]} x, \ x \in \mathbb{R}^1, \tag{2.7}$$

where $\hat{\beta}$ is the least squares estimator in the model

$$\begin{aligned} &Y_i = \beta X_i + \varepsilon_i, \ X_1, \ldots, X_n \ \mathbb{R}\text{-valued and i.i.d. with } \mathbb{E}|X_i|^2 = 1, \\ &\{\varepsilon_i\}_i \text{ i.i.d. and independent from } \{X_i\}_i, \ \mathbb{E}[\varepsilon_i] = 0, \ \mathrm{Var}(\varepsilon_i) = \sigma^2 < \infty. \end{aligned} \tag{2.8}$$

The threshold is assumed to be of the form

$$u_n = u_n(c) = c\sigma n^{-1/2}. \tag{2.9}$$

This choice leads to a stable predictor $\hat{\theta}_n(x)$ according to Definition 1.2. But instability arises when scaling the predictor with $n^{1/2}$ which becomes an interesting case for bagging.

Proposition 2.2 *Assume model (2.8) with* $\beta = \beta_n(b) = b\sigma n^{-1/2}$ *and* $\mathbb{E}|\varepsilon_i|^4 < \infty$, $\mathbb{E}|X_i|^4 < \infty$. *For the predictor in (2.7) with* $u_n = u_n(c)$ *as in (2.9),*

$$n^{1/2}\sigma^{-1}\hat{\theta}_n(x) \to_D g(Z_b) = (Z_b - Z_b \mathbb{I}_{[|Z_b| \le c]})x,$$
$$n^{1/2}\sigma^{-1}\hat{\theta}_{n;B}(x) \to_D g_B(Z_b),$$

where $Z_b = b + Z,\ Z \sim \mathcal{N}(0,1),$ *and*

$$g_B(z) = (z - \{z\Phi(c - z) - \varphi(c - z) - z\Phi(-c - z) + \varphi(-c - z)\})x.$$

A proof is given in section 6. The interpretation is similar to the one in section 2.1: the original predictor is approximated by $g(\cdot)$ which involves a hard-threshold indicator, whereas the bagged predictor by $g_B(\cdot)$ which is a soft-threshold function. The functions $g(\cdot)$ and $g_B(\cdot)$ are displayed in Figure 2.3.

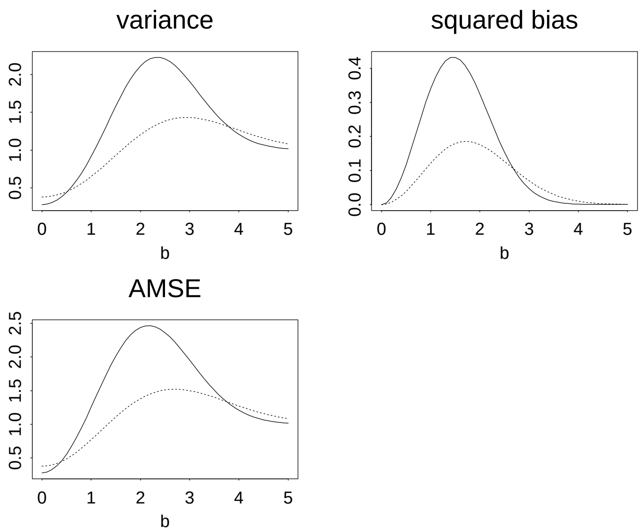

Figure 2.4: Predictor from (2.7), with $x = 1$ and u_n as in (2.9) with $c = 1.96$, in linear model (2.8) with $\beta = \beta_n(b)$ as in Proposition 2.2. Asymptotic variance, squared bias and mean squared error [AMSE], standardized by the factor $n\sigma^{-2}$, for the predictor $\hat{\theta}_n(1)$ [solid line] and for the bagged predictor $\hat{\theta}_{n;B}(1)$, as a function of b.

From Proposition 2.2 we can numerically compute the bias, variance and MSE of $\hat{\theta}_n(x)$ and $\hat{\theta}_{n;B}(x)$ as a function of b where $\beta = \beta_n(b)$ as in Proposition 2.2 [similarly as in Corollary 2.1 and using uniform integrability]. The results are displayed in Figure 2.4 using the threshold $c = \Phi^{-1}(0.975) = 1.96$ in (2.9) which arises for n large under two-sided t-testing on significance level 5%. The gain with bagging is quite substantial in the range $1 \leq b \leq 3$; for the point $b = 0$ with small amount of instability, bagging decreases performance a bit. The bias and mean squared error are here defined for estimating the true quantity $\beta x = \beta_n(b)x = b\sigma n^{-1/2}x$; this is different from the centering in section 2.1 where the original predictor is assumed to be asymptotically unbiased. In this particular setting, bagging even has smaller asymptotic bias, for most values of b; but the bias effect plays again a negligible role in terms of MSE.

2.3 MARS: a soft decision algorithm

In this section, we analyze the bagging effect on the basic ingredient in MARS [Friedman, 1991], a piecewise linear spline function. This will turn out to be a soft rather than a hard decision operation. Bagging continues to act as smoothing making the decision even softer, but this smoothing effect only brings in a marginal reduction in variance and renders bagging unnecessary for MARS.

For a one-dimensional predictor, the basis function in MARS is a piecewise linear spline function $[x - d]_+ = (x - d)\mathbb{I}_{[d \leq x]}$. Its estimated version takes the form

$$\hat{\theta}_n(x) = \hat{\beta}_n[x - \hat{d}_n]_+, \tag{2.10}$$

with the least squares estimates

$$(\hat{\beta}_n, \hat{d}_n) = \mathrm{argmin}_{\beta,d} \sum_{i=1}^{n} (Y_i - \beta[X_i - d]_+)^2 \tag{2.11}$$

for the best projected values

$$(\beta^0, d^0) = \mathrm{argmin}_{\beta,d}\, \mathbb{E}[(Y - \beta[X - d]_+)^2]. \tag{2.12}$$

These estimators behave differently from the hard decision algorithms in a crucial way so that bagging turns out to be non-effective. We illustrate it in the regression model,

$$Y_i = f(X_i) + \varepsilon_i,\ \mathrm{supp}(X_i) = \mathcal{D} \subseteq \mathbb{R}^1 \text{ an open set},\ \mathrm{supp}(\varepsilon_i) = \mathbb{R}\ \ (i = 1, \ldots, n), \tag{2.13}$$

where $\{X_i\}$ and $\{\varepsilon_i\}_i$ are i.i.d. sequences, independent of each other. Moreover, $\mathbb{E}[\varepsilon_i] = 0,\ \mathrm{Var}(\varepsilon_i) = \sigma^2 < \infty$.

Proposition 2.3 *Consider the regression model (2.13) with* $\mathbb{E}|Y_i|^2 < \infty$, $\mathbb{E}|X_i|^2 < \infty$. *Assume the density function for* X_i *is positive everywhere and bounded over a neighborhood of the best projected parameter* d^0. *Then, the estimators in (2.11) are asymptotically independent and*

$$\sqrt{n}(\hat{\beta}_n - \beta^0) \to_D \mathcal{N}(0, \sigma^2_\beta),$$
$$\sqrt{n}(\hat{d}_n - d^0) \to_D \mathcal{N}(0, \sigma^2_d),$$

where β^0, d^0 *are as in (2.12).*

Proof: The argument is essentially the same as that in Chan and Tsay (1998), noting that finite second moments are sufficient for independent data. □

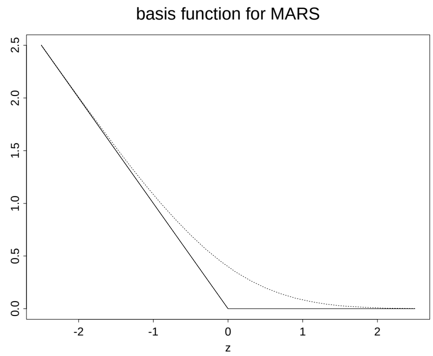

Figure 2.5: MARS basis function at $x = x_n(0) = d^0$ as in (2.14). Function $g(z)$ [solid line] and $g_B(z)$ [dotted line] from Proposition 2.4, defining the asymptotics of $\hat{\theta}_n(x_n(0))$ and its bagged version ,respectively.

Following Proposition 2.3, the MARS predictor (2.10) in the simplest case is stable in the sense of Definition 1.2, even for x in an $n^{-1/2}$-neighborhood of d^0. Note that this is not true for the indicator case in section 2.1, but it does hold for the predictor in the variable selection problem from (2.7). Due to the hard decision in the latter case, bagging brought in a substantial improvement in terms of the leading MSE of order $O(n^{-1})$ [see Proposition 2.2 and Figure 2.4].

Consider now explanatory variables which are in a region around the non-differentiable point [the 'kink'] of the MARS predictor,

$$x = x_n(c) = d^0 + c\sigma_d n^{-1/2}. \tag{2.14}$$

The smoothing effect of bagging with MARS can be described now.

Proposition 2.4 *Under the conditions of Proposition 2.3,*

$$n^{1/2}\sigma_d^{-1}\hat{\theta}_n(x_n(c)) \to_D g(Z) = \beta^0(c - Z)\mathbb{I}_{[Z \le c]},$$
$$n^{1/2}\sigma_d^{-1}\hat{\theta}_{n;B}(x_n(c)) \to_D g_B(Z) = \beta^0\{(c - Z)\Phi(c - Z) + \varphi(c - Z)\},$$

where $Z \sim \mathcal{N}(0, 1)$.

A proof is given in section 6. The functions $g(\cdot)$ and $g_B(\cdot)$ are displayed in Figure 2.5, and the MSEs displayed in Figure 2.6 are obtained by integrating the limiting quantities from Proposition 2.4 [assuming moment conditions]. In contrast, for the continuous MARS decision [see Figure 2.5], the bagging improvement is almost negligible.

The results for the basic MARS predictor (2.10) are also found to be relevant for more complex predictions with MARS in section 4. In summary, our theoretical analysis does indeed explain [partially] when bagging works: it improves very little in the case of the continuous-decision MARS procedure, but very much upon procedures involving hard, discontinuous decisions.

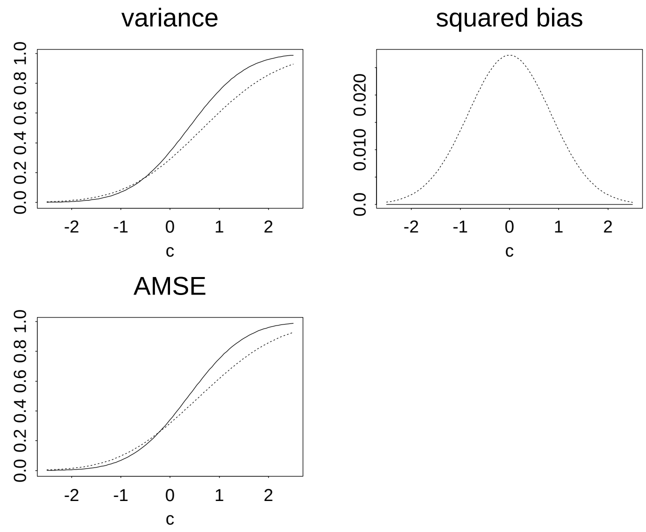

Figure 2.6: MARS predictor $\hat{\theta}_n(x_n(c))$ from (2.10) with $x_n(c)$ from (2.14). Asymptotic variance, squared bias and mean squared error [AMSE], standardized by the factor $n\sigma_d^{-2}$, for the predictor $\hat{\theta}_n(x_n(c))$ [solid line] and for the bagged predictor $\hat{\theta}_{n;B}(x_n(c))$ [dotted line], as a function of c.

3 Subagging decision trees

In this section, we address the effect of bagging in the case of decision trees, which is the most commonly used procedure for bagging in empirical studies. Decision trees consist of piecewise constant fitted functions sitting on products of indicator functions as in (2.6). Hence we expect bagging to bring significant variance reduction as in section 2.1. To make the case rigorous, we have to derive in section 3.1 a new result on the asymptotic distribution of the split point in decision trees. The rate is $n^{-1/3}$, however, and the bootstrap doesn't work for the split point. A rigorous analysis of bagging is thus very difficult. But subagging, a computationally more efficient alternative, turns out to be more tractable and we develop an upper bound on the variance of the subagged predictor based on trees. Section 3.3 contains a small simulation study to show that the asymptotic result in section 3.2 is relevant for small sample sizes. Section 3.4 makes the argument that the local smoothing effects around the split points come together to give rise to a global scale variance reduction. This is because there are a large number of split points and hence a large number of unstable regions where bagging improves over the original decision tree.

3.1 Cube-root asymptotics for the one-split stumps

For a one-dimensional predictor space, a non-normal limiting distribution is derived for the split point in stumps, i.e. a binary tree with two terminal nodes. It is the basis for our rigorous analysis of aggregation with stumps and its implications for large binary trees.

In model (2.13), consider now the decision tree predictor with stumps,

$$\hat{\theta}_n(x) = \hat{\beta}_\ell \mathbb{I}_{[x<\hat{d}_n]} + \hat{\beta}_u \mathbb{I}_{[x\geq\hat{d}_n]}, \tag{3.1}$$

where the estimates are obtained by least squares as

$$(\hat{\beta}_\ell, \hat{\beta}_u, \hat{d}_n) = \text{argmin}_{\beta_\ell,\beta_u,d} \sum_{i=1}^{n} (Y_i - \beta_\ell \mathbb{I}_{[X_i<d]} - \beta_u \mathbb{I}_{[X_i\geq d]})^2. \tag{3.2}$$

The best projected values are defined by

$$(\beta_\ell^0, \beta_u^0, d^0) = \text{argmin}_{\beta_\ell,\beta_u,d} \mathbb{E}[(Y - \beta_\ell \mathbb{I}_{[X<d]} - \beta_u \mathbb{I}_{[X\geq d]})^2]. \tag{3.3}$$

Solving the normal equations of (3.3) gives

$$\beta_\ell^0 = \mathbb{E}[Y|X < d^0],\ \beta_u^0 = \mathbb{E}[Y|X \geq d^0],\ f(d^0) = \frac{\beta_\ell^0 + \beta_u^0}{2}$$

with $f(\cdot)$ from (2.13). To proceed, we make the following assumptions for model (2.13).

(A2) (i) [smoothness condition on f] $f(\cdot)$ is continuous; and its first and second derivatives f', f'' exist and are uniformly bounded in a neighborhood of d^0 and $f'(d^0) \neq 0$.

(ii) [smoothness condition on the density functions of X and ε] X_i and ε_i have density functions p_X and p_ε respectively; the first derivative p'_X exists and is uniformly bounded in a neighborhood of d^0, and $p_X(d^0) \neq 0$.

(iii) [moment condition] $\mathbb{E}[\varepsilon_i] = 0, \mathbb{E}|\varepsilon_i|^2 = \sigma^2 < \infty$;

(iv) [tail condition] the marginal density p_Y of Y satisfies $p_Y(y) = o(|y|^{-(4+\delta)})$ as $|y| \to \infty$, for some $\delta > 0$.

Condition (iv) is satisfied, for example, when the same tail condition holds for p_ε, as in the case of Gaussian noise, and f is bounded on its domain $\mathcal{D}$.

Theorem 3.1 *Suppose assumption (A2) holds, $\beta_\ell^0 \neq \beta_u^0$, and the best projected values $(\beta_\ell^0, \beta_u^0, d^0)$ are unique. Then as $n \to \infty$,*

$$n^{1/3}(\hat{d}_n - d^0) \to_D W := \mathit{argmax}_t[Q(t)\mathit{sign}(\beta_\ell^0 - \beta_u^0)],$$

where the limiting process Q is a scaled, two-sided Brownian motion, originating from zero, with a quadratic drift:

$$Q(t) = \sigma_0 B(t) - \frac{1}{2}Vt^2,$$

where $\sigma_0^2 = p_X(d^0)\sigma^2$, $B(t)$ a two-sided Brownian motion, originating from zero, and $V = -p_X(d^0)f'(d^0) \neq 0$.

A proof is given in section 6.

Remark 3.1. Theorem 3.1 generalizes to the case where X in (2.13) is p-dimensional with $p > 1$. All what is required is that the theoretically optimal component $\gamma^0 \in \{1, \ldots, p\}$ to split is unique.

Remark 3.2. The analysis for best-first induced binary trees with finitely many splits [i.e. without pruning] is similar to Theorem 3.1. More details are given by Fact 3.1 in section 3.4.

Groeneboom (1989, corollary 3.1) studies the distribution of the maximizer of the process $B(t) - ct^2$ ($c > 0$) and gives its density function. Unfortunately, this density is not normal and involves functions whose Fourier transforms are characterized by Airy functions. Since it is in no sense simple and does not give any insights into the distribution of W, we refer interested readers to Groeneboom (1989). Thus, the asymptotic normality assumption (A1) for $\hat{d}_n$ in section 2 does not hold! Moreover, the bootstrapped estimator $\hat{d}_n^*$, when centered around $\hat{d}_n$, does not converge to the same limiting distribution as that of W. The proof of Theorem 3.1 offers some insights. The $n^{-1/3}$-asymptotics holds largely due to the smoothness conditions in (A2) on the population density and conditional density functions. These conditions are violated for the bootstrapped samples, for which the underlying distribution is discrete.

It is worth noting that (A1) about bootstrap consistency is not necessary for bagging to work as long as the resulting bagged estimator is sensible itself. Conditional on the original sample, $\hat{d}_n^*$ spreads around d^0 by taking one of the discrete values between original sample points. The resulting bagged stump estimator is a weighted average of the stump estimators with split points between the original sample values. Thus, bagging is still a smoothing operation, similar to the assertion in Proposition 2.1, although exact analysis seems difficult and we leave it as an open research problem. As a computationally more efficient alternative which is also accessible for analysis, we study next a variant of the bagging procedure.

3.2 Subagging

Subagging is a sobriquet for '**sub**sample **agg**regat**ing**' where subsampling is used instead of the bootstrap for the aggregation. A predictor $\hat{\theta}_n(x) = h_n(L_1, \ldots, L_n)(x)$ is aggregated as follows:

$$\hat{\theta}_{n;SB(m)} = \left(\begin{array}{c} n \\ m \end{array} \right)^{-1} \sum_{(i_1,\ldots,i_m)\in\mathcal{I}} h_m(L_{i_1}, \ldots, L_{i_m}), \tag{3.4}$$

where $\mathcal{I}$ is the set of m-tuples whose elements in $\{1, \ldots, n\}$ are all distinct. This aggregation can be approximated by a stochastic computation: random sampling m times of the data $L_1, \ldots, L_n$ without replacement and averaging over the predictors based on random subsamples, cf. Bickel et al. (1997).

We first consider an arbitrary predictor and then specialize to the examples in (2.6) and (3.1).

Proposition 3.1 *Let $\hat{\theta}_n(\cdot) = h_n(L_1, \ldots, L_n)(\cdot)$ be any predictor which is symmetric in the data $L_1, \ldots, L_n$. Assume that $m \le n$ and $\mathbb{E}|h_m(L_1, \ldots, L_m)(x)|^2 < \infty$ for all x. Then, for any x,*

$$\begin{aligned} \mathbb{E}[\hat{\theta}_{n;SB(m)}(x)] &= \mathbb{E}[h_m(L_1, \ldots, L_m)(x)], \\ \mathrm{Var}(\hat{\theta}_{n;SB(m)}(x)) &\le \frac{m}{n} \mathrm{Var}(h_m(L_1, \ldots, L_m)(x)). \end{aligned}$$

Proof: The subagged predictor $\hat{\theta}_{n;SB(m)}(x)$ is a U-statistic with kernel of order m. The result then follows from a well known formula for the variance of a U-statistic, cf. Serfling (1980). $\square$

3.2.1 Fraction and half subagging

An interesting case is subagging with $m = [an]$ with $0 < a < 1$ [i.e. m a fraction of n] and often $a = 1/2$ [half subagging]; and *not* with m of smaller order than n which will be discussed in section 3.2.2. The choice $a = 1/2$ is also suggested by Friedman and Hall (2000), mainly by simulations.

We assume now the following very mild condition for the predictor in (2.6) or (3.1):

(A3) For some increasing sequence $(b_n)_{n\in\mathbb{N}}$,

$$\mathbb{P}[b_n(\hat{d}_n - d^0) \leq x] \to G(x)$$

where $G(\cdot)$ is the c.d.f. of a non-degenerate distribution.

By Theorem 3.1, assumption (A3) holds for the split point in stumps with

$$b_n = n^{1/3}\sigma_\infty^{-1},\ \sigma_\infty^2 = \lim_{n\to\infty} n^{2/3}\,\mathrm{Var}(\hat{d}_n) = Var(W), \tag{3.5}$$

where W is as in Theorem 3.1. We evaluate expectation and variance of subagged estimators for the predictors in (2.6) and (3.1) at unstable locations. In the case of stumps (3.1), the explanatory variable x is in an $n^{-1/3}$-neighborhood of d^0,

$$x = x_n(c) = d^0 + c\sigma_\infty n^{-1/3}, \tag{3.6}$$

with σ_∞^2 from (3.5).

Theorem 3.2 *[Fraction subagging for indicators and stumps]*
Consider predictors as in (2.6) or (3.1) with $x = x_n(c)$ as in Proposition 2.1 or (3.6), respectively. Assume that (A3) holds. Suppose $m = [an]$ with $0 < a < 1$. Then,

$$\begin{aligned}
&\lim_{n\to\infty}\mathbb{E}[\hat{\theta}_{n;SB(m)}(x_n(c))] = \beta_\ell^0 + (\beta_u^0 - \beta_\ell^0)G(cb_m/b_n),\\
&\limsup_{n\to\infty}\mathrm{Var}(\hat{\theta}_{n;SB(m)}(x_n(c))) \leq (\beta_u^0 - \beta_\ell^0)^2 aG(cb_m/b_n)(1 - G(cb_m/b_n))\\
&\limsup_{n\to\infty}\mathbb{E}[\Big(\hat{\theta}_{n;SB(m)}(x_n(c)) - \mathbb{E}[\hat{\theta}_n(x(c))]\Big)^2]\\
&\leq (\beta_u^0 - \beta_\ell^0)^2\left((G(cb_m/b_n) - G(c))^2 + aG(cb_m/b_n)(1 - G(cb_m/b_n))\right),
\end{aligned}$$

where $\beta_\ell^0 = 0,\ \beta_u^0 = 1$ for the predictor in (2.6).

A proof is given in section 6. The evaluation of the asymptotic MSE [AMSE] bounds in Theorem 3.2 depends on the normalizing constants b_n and the limiting distribution $G(\cdot)$ in (A3). If $b_n = C\sqrt{n}$ [C a constant] and $G(\cdot) = \Phi(\cdot)$ the standard Gaussian c.d.f., the evaluation is straightforward and the result is displayed in the top panel of Figure 3.1. In the case of the stumps predictor, we know that $b_n = Cn^{1/3}$ [C a constant] and $G(\cdot)$ can be characterized in terms of Airy functions: a more explicit form for $G(\cdot)$ is not possible. We thus rely on simulating the asymptotic distribution $G(\cdot)$ and display the result in the bottom panel of Figure 3.1. The description of subagging with larger decision trees is postponed to section 3.4.

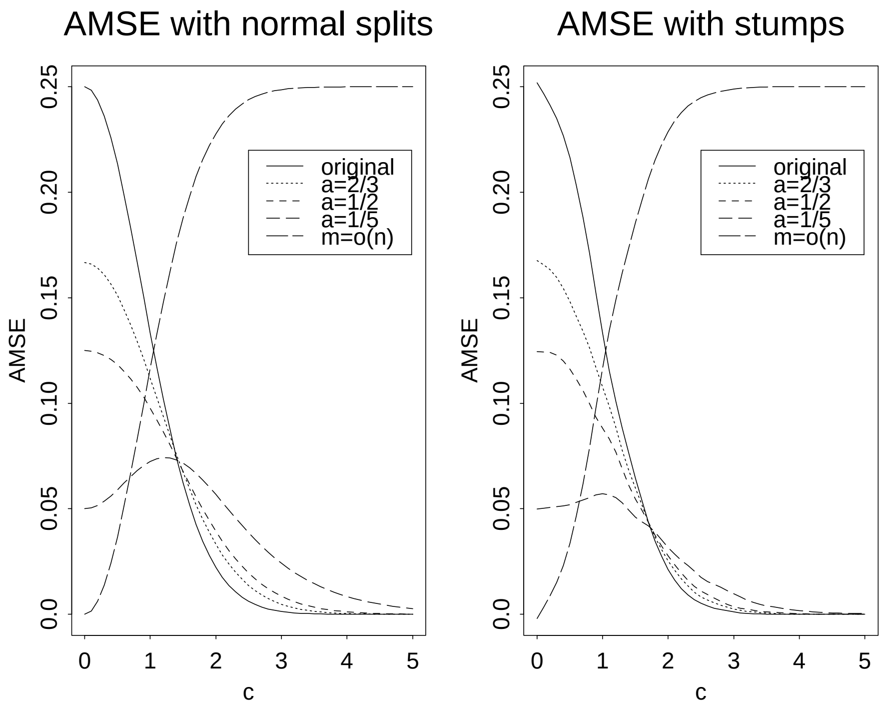

Figure 3.1: Asymptotic mean squared error [AMSE] of original predictor and a bound for the subagged version. Top: indicator predictor $\hat{\theta}_n(x_n(c))$ in (2.6) [solid line] and $\hat{\theta}_{n;SB(m)}(x_n(c))$, with $x_n(c)$ as in Proposition 2.1. The situation corresponds to Theorems 3.2 and 3.3, assuming (A3) with $b_n = n^{1/2}\sigma_\infty^{-1}$ and $G(\cdot) = \Phi(\cdot)$ the standard normal distribution. Bottom: $\hat{\theta}_n(x_n(c))$ in (3.1) [solid line] and $\hat{\theta}_{n;SB(m)}(x_n(c))$, with $x_n(c)$ as in (3.6). The situation corresponds to Theorems 3.2 and 3.3, assuming (A3) with $b_n = n^{1/3}\sigma_\infty^{-1}$ and $G(\cdot)$ from Theorem 3.1. In both cases: subsample size $m = [an]$ or $m \to \infty,\ m = o(n)$. Everything scaled to $\beta_\ell^0 = 0,\ \beta_u^0 = 1$.

3.2.2 Small order subagging

We refer to small order subagging when using a subsample size $m = m(n)$ so that $m \to \infty,\ m = o(n)$. This is a classical approach with subsampling for distribution estimation, cf. Bickel et al. (1997). However, such a choice is not very appropriate for subagging, as explained in the next Theorem.

Theorem 3.3 *[Small order subagging for indicators and stumps]*
Assume the same conditions as in Theorem 3.2 but with $m \to \infty,\ m = o(n)$. Then,

$$\lim_{n\to\infty} \mathbb{E}[\hat{\theta}_{n;SB(m)}(x_n(c))] = \beta_\ell^0 + (\beta_u^0 - \beta_\ell^0)G(0),$$
$$\lim_{n\to\infty} \mathrm{Var}(\hat{\theta}_{n;SB(m)}(x_n(c))) = 0,$$
$$\lim_{n\to\infty} \mathbb{E}[(\hat{\theta}_{n;SB(m)}(x_n(c)) - \mathbb{E}[\hat{\theta}_n(x(c))])^2] = (\beta_u^0 - \beta_\ell^0)^2(G(0) - G(c))^2,$$

where $\beta_\ell^0 = 0,\ \beta_u^0 = 1$ for the predictor in (2.6).

Proof: The results follow as for Theorem 3.2 by noting that $m/n = o(1)$ [which plays the role of a in Theorem 3.2] and $b_m/b_n = o(1)$ since $b_n = Cn^{1/3}$. $\square$

Numerical evaluation for small order subagging is also displayed in Figure 3.1. In the very regular case corresponding to the top panel in Figure 3.1, fraction subagging with $a = 1/5$ can already become quite bad for 'weak unstable regions' where $1.5 \le |c| \le 4.5$. The situation is contrasted somewhat with stumps displayed in the bottom panel of Figure 3.1: fraction subagging with $a = 1/5$ is not behaving poorly at 'weak unstable regions' but improves very much at 'strong unstable regions' with $|c|$ small [the latter is also true in the top panel of Figure 3.1]. Small order subagging with $m = o(n)$ can be very bad at 'weak unstable regions', in both cases corresponding to Figure 3.1. All this should be cautiously interpreted because we give only an upper bound for the AMSE in fraction subagging and actual performance may be better than this bound. Generally, the subsample size m can be interpreted as a 'smoothing' parameter: m large corresponds to small bias but large variance, and vice versa. From this view, small order subagging oversmoothes and hence magnifies the bias.

3.3 Discussion

All the quantifications in the previous sections hold in the limit. But Table 3.1 and Figure 3.3 show finite-sample situations for stumps $\hat{\theta}_n(x)$ with $n = 100$ and $n = 10$ in the model (2.13) with $f(x) = 2 + 3x$, $X_i \sim \text{Uniform}([0, 1])$ and $\varepsilon_i \sim \mathcal{N}(0, 1)$; similarly as before, the centering for bias and mean squared error is always around $\theta_n(x) = \mathbb{E}[\hat{\theta}_n(x)]$. Bagging and

n	unbagged [MSE]	bagging [MSE]	half subagging [MSE]
100	0.076	0.033 (56%)	0.031 (59%)
10	0.244	0.172 (30%)	0.170 (30%)

Table 3.1: Overall mean squared error $\mathbb{E}[(\hat{\theta}_n(X) - \theta_n(X))^2]$ [with X independent from the training data] for stumps $\hat{\theta}_n(\cdot)$ in (3.1) and its bagged and subagged ($m = n/2$) version. Reduction with (su-)bagging is given in parentheses. Based on 100 simulations from model (2.13).

half subagging are almost identical; a fact which we re-discover again in more complex situations in section 4. The reduction in MSE with (su-)bagging is larger for $n = 100$ than for $n = 10$, but still substantial for the small sample size. The result with $n = 10$ is quite important because with a deep split in a decision tree such as CART, only such a small number of observations may belong to the partition cell to be refined. Figure 3.2 with $n = 100$ is qualitatively like the asymptotic situation in Figure 3.1 [bottom panel]. There is a quantitative difference due to the fact that Figure 3.1 only shows bounds for the asymptotic mean squared error which might be too conservative. For the case here, we get a bound of about 50% on the MSE reduction around the most unstable point $c = 0$ while the actual reduction is 59%.

Our theoretical analysis and its numerical illustrations have only been dealing with a somewhat limited notion of bias. We have usually given an a priori advantage to the unbagged predictor and considered performance for estimating $\theta(x) = \lim_{n\to\infty} \mathbb{E}[\hat{\theta}_n(x)]$.

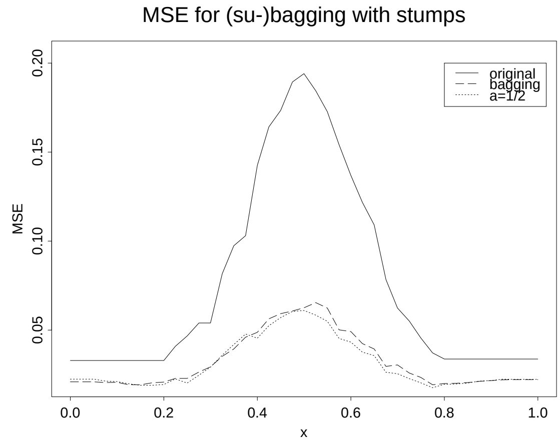

Figure 3.2: Mean squared error of stumps $\hat{\theta}_n(x)$ in (3.1) [solid line] and its (su-)bagged version $\hat{\theta}_{n;SB(m)}(x)$ for $x \in [0,1]$. Sample size $n = 100$ and subsampling size $m = [an]$. Everything multiplied by the factor $1/(\beta_u^0 - \beta_\ell^0)^2 = 1/2.25$ to obtain [asymptotically] the scale from Figure 3.1.

(Su-)bagging adds a small bias from this view. With subagging, the bias decreases as $m \leq n$ increases. More finite sample results about bias are given in section 4.

As an alternative to subagging, we briefly point to moon-bagging, standing for '**m o**ut **of n b**ootstrap **agg**regat**ing**'. The idea is to replace the bootstrap step by the m out of n bootstrap [Bickel et al., 1997]: sample with replacement

$$L_1^*, \dots, L_m^* \; i.i.d. \sim \hat{F}_n, \tag{3.7}$$

where $\hat{F}_n$ is the empirical distribution of the data $L_1, \dots, L_n$ and m is an integer smaller than sample size n. Then, calculate

$$\hat{\theta}_m^*(x) = h_m(L_1^*, \dots, L_m^*)(x)$$

where $\hat{\theta}_n(x) = h_n(L_1, \dots, L_n)(x)$. The moon-bagged predictor with resampling size m is then

$$\hat{\theta}_{n;MB(m)}(x) = \mathbb{E}^*[\hat{\theta}_m^*(x)].$$

The difference between moon-bagging and subagging essentially disappears for m small [with respect to n]; particularly, Theorem 3.3 also applies for moon-bagging [if $h_n(\cdot)(x)$ is not greatly affected by ties], cf. Bickel et al. (1997).

3.4 Global mean squared error and trees with many terminal nodes

This section discusses the relevance of our results about (su-)bagging with stumps to a general binary decision tree with many terminal nodes and predictor space $\mathbb{R}^p$ with $p > 1$.

3.4.1 Stumps with one split

First we use Theorem 3.1 to assess the effect of subagging on the global mean squared error for the one-split stumps. Recall that $\theta(x) = \lim_n \mathbb{E}[\hat{\theta}_n(x)]$ has been defined as the asymptotic value of the original predictor which is a suitable target when comparing the original with the (su-)bagged procedure, because

$$\mathbb{E}[(\hat{\theta}_n(x) - f(x))^2] \sim \mathbb{E}[(\hat{\theta}_n(x) - \theta(x))^2] + (\theta(x) - f(x))^2,$$

where the last term will not be affected by the (su-)bagging aggregation. Denote by

$$\mathrm{MSE}_n = \mathbb{E}[(\hat{\theta}_n(X) - \theta(X))^2]$$

for a new test observation $X \in \mathbb{R}$ [notationally simpler than $\mathbb{R}^p$] which is independent from the data, having the same distribution as one predictor in the data. Denoting by $p_X(\cdot)$ the density for X, we rewrite

$$\mathrm{MSE}_n = \int \mathrm{MSE}_n(x) p_X(x) dx,$$

where $\mathrm{MSE}_n(x) = \mathbb{E}[(\hat{\theta}_n(x) - \theta(x))^2]$ for fixed x. With one split [stumps], the instability region is in a $n^{-1/3}$-neighborhood of the best projected value d^0. Rewrite by setting $x = d^0 + vn^{-1/3}$,

$$\mathrm{MSE}_n = n^{-1/3} \int \mathrm{MSE}_n(d^0 + vn^{-1/3}) p_X(d^0 + vn^{-1/3})) dv.$$

Assuming that $p_X(\cdot)$ is continuous in a neighborhood of d^0 we have $p_X(d^0 + vn^{-1/3})) \to p_X(d^0)$. Moreover, Theorem 3.1 indicates that $\mathrm{MSE}_n(d^0 + vn^{-1/3}) \to m(v)$ for some function $m(\cdot): \mathbb{R} \to \mathbb{R}^+$. Assuming regularity conditions to interchange the integration with the limiting operation [e.g. for applying Lebesgue's Dominated Convergence Theorem], we get

$$\mathrm{MSE}_n \sim n^{-1/3} p_X(d^0) \int_{-\infty}^{\infty} m(v) dv.$$

Analogously, we obtain for the (su-bagged) predictor, but now with a different function $m_{SB}(\cdot)$,

$$\mathrm{MSE}_{n;SB} \sim n^{-1/3} p_X(d^0) \int_{-\infty}^{\infty} m_{SB}(v) dv.$$

Our rigorous analysis in subsection 3.2 has shown that $m_{SB}(v) << m(v)$ for v close to zero, and $m_{SB}(v) < m(v)$ for most $v \in \mathbb{R}$ with $m(v),\ m_{SB}(v)$ not very close to zero. We thus conclude that the gain with (su-)bagging for stumps is asymptotically given by

$$\mathrm{MSE}_{n;SB}/\mathrm{MSE}_n \sim \int m_{SB}(v) dv \Big/ \int m(v) dv, \tag{3.8}$$

which is usually much smaller than one. Using Remark 3.1 and the same arguments from above, this easily generalizes to stumps with p-dimensional covariate space.

3.4.2 General binary decision trees

Let us first consider a two split [three terminal node] decision tree in the case of a one-dimensional predictor space as a generalization to the stumps result in Theorem 3.1. The first split $\hat{d}_1$ is estimated as with stumps in (3.2), leading to two partition cells $\mathcal{R}_\ell = \{x : x < \hat{d}_1\}$ and $\mathcal{R}_u = \{x : x \geq \hat{d}_1\}$. Without loss of generality assume that asymptotically, the lower partition cell $\mathcal{R}_\ell$ will be refined with a second split $\hat{d}_2$, defined as

$$(\hat{\beta}_{2;\ell}, \hat{\beta}_{2;u}, \hat{d}_2) = \mathrm{argmin}_{\beta_{2;\ell},\beta_{2;u},d_2<\hat{d}_1} \sum_{i=1}^{n}(Y_i - (\beta_{2;\ell}\mathbb{I}_{[X_i<d_2]} + \beta_{2;u}\mathbb{I}_{[X_i\geq d_2]} + \hat{\beta}_{1;u}\mathbb{I}_{[X_i\geq \hat{d}_1]}))^2,$$

where $\hat{\beta}_{1;u}$ is the estimated location for the upper partition cell $\mathcal{R}_u$ from the first split. The following can then be shown.

Fact 3.1 *Under similar conditions as in Theorem 3.1, but now for the conditional densities of $Y|X < d_1^0$ and $X|X < d_1^0$ [where d_1^0 is the best projected value for the first split as in (3.3)],*

$$n^{1/3}(\hat{d}_2 - d_2^0) \to_{\mathcal{D}} W_2,$$

where W_2 is a maximizer of a two-sided Brownian motion with quadratic drift, similar to Theorem 3.1, and d_2^0 the best projected second split.

A sketch of a proof is given in section 6. Note that the second split has the same convergence rate $n^{-1/3}$ but the limiting distribution of W_2 might have a different scale [variance] from the one from the first split described in Theorem 3.1. Nevertheless, (su-)bagging has about the same relative variance reduction effect on the second as on the first split.

Consider now the global MSE with two splits and optimal projected values for the first and second split d_1^0 and d_2^0, respectively [without loss of generality assume $d_1^0 > d_2^0$]. Then, we write

$$\mathrm{MSE}_n = \int_{-\infty}^{d_2^0+\kappa} \mathrm{MSE}_n(x)p_X(x)dx + \int_{d_2^0+\kappa}^{\infty} \mathrm{MSE}_n(x')p_X(x')dx',$$

where $\kappa > 0$ is arbitrary small. Now use the substitutions $x = d_2^0 + vn^{-1/3}$ and $x' = d_1^0 + v'n^{-1/3}$. Due to Theorem 3.1 and Fact 3.1, $\mathrm{MSE}_n(x)$ and $\mathrm{MSE}_n(x')$ converge to $m_2(v)$ and $m_1(v')$, respectively. Assume that regularity conditions to interchange integration with the limiting operation hold, as in the case with stumps. Then, for a two split tree,

$$\mathrm{MSE}_n \sim n^{-1/3}[p_X(d_1^0)\int m_1(v)dv + p_X(d_2^0)\int m_2(v)dv].$$

Using the same arguments for (su-)bagging suggests

$$\mathrm{MSE}_{n;SB} \sim n^{-1/3}[p_X(d_1^0)\int m_{1;SB}(v)dv + p_X(d_2^0)\int m_{2;SB}(v)dv].$$

Now, Theorem 3.2 [use also Fact 3.1 for the seconds split] suggests a reduction so that both

$$\int m_{i;SB}(v)dv / \int m_i(v)dv << 1,\ i = 1, 2. \tag{3.9}$$

For a two split tree, elementary algebra then leads to

$$\limsup_{n\to\infty}(\mathrm{MSE}_{n;SB}/\mathrm{MSE}_n) \le \max_{i=1,2}\left(\int m_{i;SB}(v)dv/\int m_i(v)dv\right),$$

which is substantially smaller than one due to (3.9). The relative gain with (su-)bagging is thus at least as big as the one for stumps in (3.8).

Fact 3.1 and the arguments about the MSE easily extend to a finite number of splits and even to the case where the number of splits grows slowly. Moreover, the argument carries over to high-dimensional covariate space (with fixed dimension $1 \le p < \infty$) and thus to the case where decision trees are most popular, see also Remark 3.1.

Since every split in a decision tree induces a whole region in the covariate space of volume $O(n^{-1/3})$ where the predictor becomes unstable, a large tree is more unstable than a small one. Instability of hard threshold decision trees has been exploited from a different view also by Loh and Shih (1997).

3.5 Subagging in classification

Aggregation in classification is often empirically found to improve similarly as in regression, see also section 4.

Consider the J-class problem: the data consists of the pairs $L_i = (Y_i, X_i)$ $(i = 1, \ldots, n)$ but now with categorical responses $Y_i \in \{0, \ldots, J-1\}$ and explanatory variables $X_i \in \mathbb{R}^p$. The task is to classify a new test variable Y based on its corresponding explanatory X: (Y, X) is independent from the data and has the same distribution as one data pair. We wish to minimize the following misclassification risk for a classifier $\mathcal{C}(\cdot)$,

$$\mathrm{MCR} = \mathbb{P}[\mathcal{C}(X) \neq Y],$$

assuming equal misclassification costs. The classifier is chosen to be of the form [as an estimated version of the optimal Bayes classifier],

$$\hat{\mathcal{C}}_n(x) = \mathrm{argmax}_j \hat{P}_n(j|x) \tag{3.10}$$

where $\hat{P}_n(j|x)$ is an estimate of $P(j|x) = \mathbb{P}[Y = j|X = x]$. (Su-)bagging of the classifier can be constructed by voting [Breiman, 1996a] or as another version [cf. Amit and Geman, 1997]

$$\hat{\mathcal{C}}_{n;SB(m)}(x) = \mathrm{argmax}_j \hat{P}_{n;SB(m)}(j|x)$$

with $\hat{P}_{n;SB(m)}(j|\cdot)$ the average of subsample-estimates, as in the regression case; and analogously for bagging instead of subagging.

A rigorous analysis comparing $\hat{\mathcal{C}}_{n;SB(m)}(\cdot)$ with $\hat{\mathcal{C}}_n(\cdot)$ when $\hat{P}_n(\cdot)$ is from a classification or decision tree is more difficult than showing the improvement with subagging in terms of MSE as given in sections 3.2-3.4. The reason is that the misclassification rate MCR$(\cdot)$ involves the distribution of $\hat{P}_n(\cdot)$ and $\hat{P}_{n;SB(m)}(\cdot)$ and not just the first two moments.

4 Numerical examples

We reconsider the two examples from Breiman (1996a) by reporting here additionally on bias and variance. Subagging as a variant of bagging is also investigate. The original predictors are either decision trees as implemented in S-Plus with the function **tree**, or MARS as implemented with the function **mars** from the library MDA in S-Plus, available from the internet at 'http://lib.stat.cmu.edu/S/mda'.

4.1 Regression setting

We consider a simulation model, called Friedman #1 (Friedman, 1991):

$$Y_i = f(X_i) + \varepsilon_i \ (i = 1, \ldots, n),$$
$$X_1, \ldots, X_n \text{ i.i.d. } \sim \text{ Uniform}_{10}([0,1]^{10}),\ \varepsilon_1, \ldots, \varepsilon_n \text{ i.i.d. } \sim \mathcal{N}(0,1),$$

where $\{X_i\}_i,\ \{\varepsilon_i\}_i$ are independent from each other, and $\text{Uniform}_p([0,1]^p)$ is given by p i.i.d. univariate Uniform([0,1]) random variables. The regression function is

$$f(x) = 10\sin(\pi x^{(1)} x^{(2)}) + 20(x^{(3)} - 1/2)^2 + 10x^{(4)} + 5x^{(5)},$$

so that the other coordinates 6 to 10 of x are not contributing to $f(\cdot)$. Sample size is chosen as $n = 500$. Our analysis is based on 100 simulation runs over the model; aggregation is computed using 50 replicates [for each model realization]. Figure 4.1 displays the results: the bias is here defined in the usual sense, namely for the true quantity $f(\cdot)$ [instead of $\theta(\cdot) = \lim_n \mathbb{E}[\hat{\theta}(\cdot)]$]. Note the different scales for decision trees and MARS. (Su-)bagging works well for trees, whereas the original MARS is already close to optimal [noise variance is 1] and (su-)bagging doesn't really improve, being consistent with the analysis of bagging in section 2.3.

We consider next the ozone data set [Breiman, 1996a]: it consists of 330 measurements of maximum daily ozone in the Los Angeles area, and 8 meteorological predictor variables. Aggregation is here computed using 25 replicates; and the mean squared error is estimated as in Breiman (1996a): random division in 90% training and 10% test set, then calculating the L^2 test set error and finally averaging those over 50 training-test-set random divisions. Figure 4.2 displays the results. (Su-)bagging works well for decision trees, whereas it yields no improvements for MARS; the (su-)bagged tree is about as good as the original [or (su-)bagged] MARS predictor.

Note that in both examples, the MSE reduction with (su-)bagging is not quite as large as in Table 3.1: this is due to the fact that the size of the bias [when centering around the true value] somewhat decreases the relative performance gain.

4.2 Classification

We consider here also the classification problem for the real data example about glass types [Breiman, 1996a]: there are 6 classes and 9 chemical measurements as predictor variables. Sample size is $n = 214$. The misclassification rate $\mathbb{P}[\mathcal{C}(X) \neq Y]$ [equal misclassification costs] is estimated with random division into training- and test-sets, analogously as for the MSE with the ozone data set in the previous section. Figure 4.3 displays the results. Bagging is slightly better than half subagging. This is one of the examples showing among

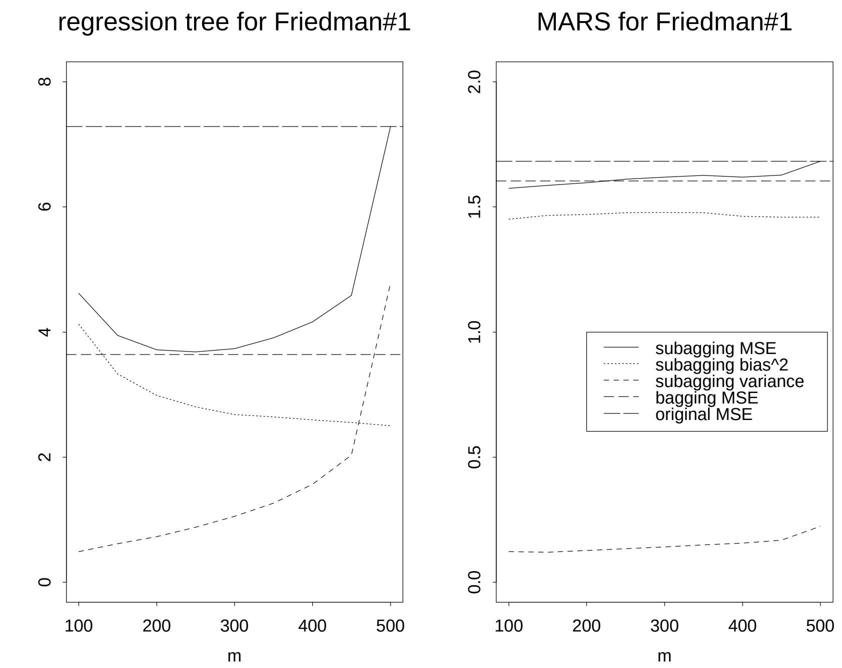

Figure 4.1: Performance for a large regression tree and MARS and their (su-)bagged versions in the simulated model Friedman #1.

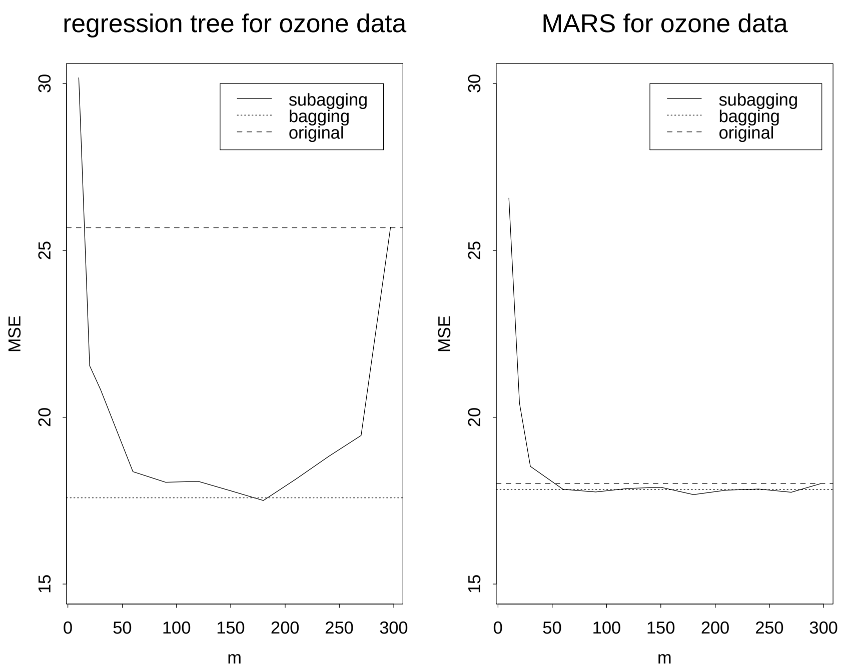

Figure 4.2: Mean squared error performance for a large regression tree and MARS and their (su-)bagged versions for the ozone data.

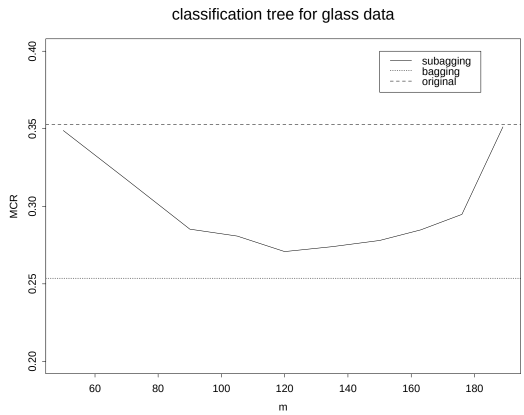

Figure 4.3: Misclassification rate [MCR] for the classifier in (3.10) with $\hat{P}_n(\cdot)$ a large tree and its (su-)bagged version.

the worst [but still small] magnitude of loss with subagging compared to bagging: relatively large subsample sizes are needed for good performance, maybe due to the relatively small training sample size 189 for trees with many splits.

5 Conclusions

We have given new theoretical arguments to explain why bagging and its variant subagging work asymptotically: they rely on the fact that the predictor is unstable in the sense of Definition 1.2. Generally, (su-)bagging doesn't make the predictor stable, but it stabilizes to a certain extent. In cases where instability comes in through hard decision indicators, arising often in many modeling techniques, (su-)bagging smoothes out the hard-thresholds yielding a soft decision scheme which lowers variance and mean squared error. Our analysis also gives more insights to the combined procedure with bagging and boosting [Bühlmann and Yu, 2000] which is very competitive. In particular, non-standard asymptotic results about stumps are given upon which we build our explanation on how (su-)bagging works for and improves upon decision trees with many terminal nodes. The theoretical results are augmented by a small simulation study with finite sample sizes to show that asymptotics kicks in rather quickly.

Moreover, we establish the fact that (su-)bagging also works for low-dimensional predictors such as stumps. This has not been greatly recognized before: for example, Breiman (1996a) and Dietterich (1996) [in his second implication] exclusively mention high dimensional schemes. Only Buja and Stuetzle (2000b) also point out that bagging is a smoothing operation, also in low-dimensional settings. We show that half subagging is as accurate as bagging but computationally cheaper. The latter is interesting for very large data sets, where fraction subagging with $m = [an]$, $a > 0$ requires much less computations but still maintains good performance [due to the fact that m has still reasonable size]. In addition, we discuss why (su-)bagging can be less effective for predictors such as MARS involving continuous decisions. This provides a partial answer to the fifth implication in Dietterich (1996), which poses the question about 'the degree of instability', or in other words the degree of improvement with (su-)bagging.

Lastly, Freund and Schapire [1998, sec. 1] raise the issue about randomness for aggregation in bagging in contrast to boosting [by deterministic reweighting]. (Su-)bagging, at least as defined theoretically, doesn't use extra randomness in the procedure. The aggregates, namely the bootstrap expectation $\mathbb{E}^*[\cdot]$ for bagging, or summing over the set $\mathcal{I}$ in (3.4) in subagging are just fixed functions of the data: but the practical *computation* is implemented by Monte Carlo. We believe that this random Monte Carlo approximation has a negligible effect on the whole problem [which is the usual view in bootstrapping or subsampling].

6 Proofs

Since the proof for Theorem 3.1 is long, we leave it to the end. Other proofs are given in order.

Proof of Proposition 2.2:
Since $\beta = 0$, $\hat{\beta} \to_D \mathcal{N}(0, \sigma^2)$. Then, by the Continuous Mapping Theorem,

$$n^{1/2}\sigma^{-1}\hat{\theta}_n(x) = g(n^{1/2}\sigma^{-1}\hat{\beta}) \to_D g(X),$$

because the set of discontinuity points of $g(\cdot)$ has Lebesgue measure zero. This proves the first assertion.

For the bagged predictor we use that

$$\sup_{v\in\mathbb{R}} |\mathbb{P}^*[n^{1/2}(\hat{\beta}^* - \hat{\beta}) \le v] - \Phi(v/\sigma)| = o_P(1),$$

cf. Freedman (1981): or in other words, $n^{1/2}(\hat{\beta}^* - \hat{\beta}) \to_D \mathcal{N}(0, \sigma^2)$ in probability. Therefore, using uniform integrability in probability for $\hat{\beta}^*$ [which is ensured by $\mathbb{E}^*|\hat{\beta}^*|^2 = O_P(1)$],

$$n^{1/2}\sigma^{-1}\hat{\theta}_{n;B}(x) \to_D \mathbb{E}_W[W\mathbb{1}_{[|W|>c]}|Z]x, \qquad (6.1)$$

where $W \sim \mathcal{N}(Z, 1),\ Z \sim \mathcal{N}(0, 1)$. The right hand side of (6.1) is

$$\mathbb{E}_W[W\mathbb{1}_{[|W|>c]}|Z]x = (Z - (\mathbb{E}_W[W\mathbb{1}_{[W\le c]}|Z] - \mathbb{E}_W[W\mathbb{1}_{[W<-c]}|Z]))x. \qquad (6.2)$$

Now, for any $v \in \mathbb{R}$,

$$\begin{aligned}\mathbb{E}_W[W\mathbb{1}_{[W\le v]}|Z] &= \int_{-\infty}^{v} w\varphi(w - Z)dw = \int_{-\infty}^{v-Z} (Z + s)\varphi(s)ds \\ &= Z\Phi(v - Z) + \int_{-\infty}^{v-Z} w\varphi(s)ds = Z\Phi(v - Z) - \varphi(v - Z). \qquad (6.3)\end{aligned}$$

Using (6.3) with $v = c$ and $v = -c$ for (6.2), we complete the proof by (6.1). □

Proof of Proposition 2.4:
We can represent the estimator as

$$\hat{\theta}_n(x_n(c)) = \beta^0(x_n(c) - \hat{d})_+ + O_P(n^{-1}), \qquad (6.4)$$

due to the convergence properties of $\hat{\beta}, \hat{d}$ and the neighborhood definition of $x_n(c)$ in (2.14). The first assertion is then immediate from (6.4) and Proposition 2.3.

For the second assertion we first show that the bootstrap works:

$$\sup_{v\in\mathbb{R}} |\mathbb{P}^*[\sqrt{n}(\hat{\beta}_n - \beta^0) \le v] - \Phi(v/\sigma_\beta)| = o_P(1),$$
$$\sup_{v\in\mathbb{R}} |\mathbb{P}^*[\sqrt{n}(\hat{d}_n - d^0) \le v] - \Phi(v/\sigma_d)| = o_P(1), \tag{6.5}$$

with σ_β^2, σ_d^2 from Proposition 2.3. We sketch an outline: the bootstrap works here for empirical processes needed to deal with the problem, cf. Giné and Zinn (1990). Then, the proof for Proposition 2.3 can be adapted for the bootstrap and (6.5) follows.

The second assertion of the Proposition follows using (6.5) and analogously to the proof of Proposition 2.2 in section 6; in particular, we use again formula (6.3). □

Proof of Theorem 3.3:
According to (3.4),

$$\mathbb{E}[\hat{\theta}_{n;SB(m)}(x_n(c))] = \mathbb{E}[h_m(L_1, \ldots, L_m)(x_n(c))],$$

and the first assertion follows by the definition of $x_n(c)$ in (3.6).

For the variance, we invoke the bound in Proposition 3.1 and use straightforward calculation as with the expected value, but now for $\mathrm{Var}(h_m(L_1, \ldots, L_m)(x_n(c)))$. □

Now we turn to the *proof of Theorem 3.1:*
Recall the definition of $(\hat{\beta}_\ell, \hat{\beta}_u, \hat{d}_n)$ in (3.2). Under weak conditions [implied by (A2)], $\hat{\beta}_\ell, \hat{\beta}_u$ converge at the conventional $n^{-1/2}$-rate to the projected values β_ℓ^0 and β_u^0 defined in (3.3). Without loss of generality, we concentrate on the limiting distribution of $\hat{d}_n$ when β_ℓ and β_u take the projected values β_ℓ^0 and β_u^0 in (3.3). That is, we consider in the sequel

$$\hat{d}_n = \mathrm{argmin}_d \sum_{i=1}^{n} (Y_i - \beta_\ell^0 \mathbb{I}_{[X_i<d]} - \beta_u^0 \mathbb{I}_{[X_i \ge d]})^2.$$

Rewrite

$$\begin{aligned} & (Y_i - \beta_\ell^0 \mathbb{I}_{[X_i<d]} - \beta_u^0 \mathbb{I}_{[X_i \ge d]})^2 - Y_i^2 \\ = \; & (\beta_\ell^0)^2 \mathbb{I}_{[X_i<d]} + (\beta_u^0)^2 \mathbb{I}_{[X_i \ge d]} - 2Y_i \beta_\ell^0 \mathbb{I}_{[X_i<d]} - 2Y_i \beta_u^0 \mathbb{I}_{[X_i \ge d]} \\ = \; & (\beta_\ell^0 - \beta_u^0)[(\beta_\ell^0 + \beta_u^0) - 2Y_i]\mathbb{I}_{[X_i<d]} + \beta_u^0(\beta_u^0 - 2Y_i). \end{aligned}$$

Assume now $\beta_\ell^0 > \beta_u^0$ [the other case $\beta_\ell^0 < \beta_u^0$ is analogous]. It follows that

$$\hat{d}_n = \mathrm{argmax}_d \sum_{i=1}^{n} g(L_i, d), \quad g(L_i, d) = (Y_i - \frac{\beta_\ell^0 + \beta_u^0}{2})\mathbb{I}_{[X_i<d]}. \tag{6.6}$$

In general, let $\{g(\cdot, \theta) : \theta \in \Theta\}$ be a class of functions indexed by a subset Θ in $\mathbb{R}^k$. Its envelope function $G_R(\cdot)$ is defined as the supremum of $g(\cdot, \theta)$ over the class

$$\mathcal{G}_R = \{|g(\cdot, \theta)| : |\theta - \theta_0| \le R\}, \; R > 0.$$

We will apply the main theorem in Kim and Pollard (1990) which gives a cube-root asymptotic limiting distribution of the maximizer of

$$P_n g(\cdot, \theta) := \frac{1}{n} \sum_{i=1}^{n} g(\xi_i, \theta)$$

where $\{\xi_i\}_i$ is a sequence of i.i.d. observations from a distribution P.

Theorem 6.1 *[Kim and Pollard, 1990].*
Let $\{\theta_n\}$ be a sequence of estimators. Suppose

(i) $P_n g(\cdot, \theta_n) \geq sup_{\theta\in\Theta} P_n g(\cdot,\theta) - o_P(n^{-2/3})$;

(ii) θ_n *converges in probability to the unique* θ_0 *that maximizes* $Pg(\cdot,\theta) = E_P g(\cdot,\theta)$;

(iii) θ_0 *is an interior point of* Θ.

Let the functions be standardized so that $g(\cdot,\theta_0) \equiv 0$. *Suppose the classes* $\mathcal{G}_R$ *for* R *near 0 are uniformly manageable for the envelopes* G_R *and satisfy*

(iv) $Pg(\cdot,\theta)$ *is twice differentiable with second derivatives matrix* $-V$ *at* θ_0;

(v) $H(s,t) = \lim_{\alpha\to\infty} \alpha P g(\cdot,\theta_0 + t/\alpha) g(\cdot, \theta_0 + s/\alpha)$ *exists for each* s, t *in* $\mathbb{R}^k$ *and* $\lim_{\alpha\to\infty} \alpha P g(\cdot,\theta_0 + t/\alpha)^2 \{|g(\cdot,\theta_0+t/\alpha)| > \varepsilon\alpha\} = 0$ *for each* $\varepsilon > 0$ *and* t *in* $\mathbb{R}^k$;

(vi) $PG_R^2 = O(R)$ *as* $R \to 0$ *and for each* $\varepsilon > 0$ *there is a constant* K *such that* $PG_R^2 \mathbb{I}_{[G_R > K]} \leq \varepsilon R$ *for* R *near 0;*

(vii) $P|g(\cdot,\theta_1) - g(\cdot,\theta_2)| = O(|\theta_1 - \theta_2|)$ *near* θ_0.

Then, the process $n^{2/3} P_n g(\cdot, \theta_0 + tn^{-1/3})$ *converges in distribution to a Gaussian process* $Q(t)$ *with continuous sample paths, expected value* $-\frac{1}{2}t'Vt$ *and covariance kernel* H. *If* V *is positive definite and if* Q *has nondegenerate increments, then* $n^{1/3}(\theta_n - \theta_0)$ *converges in distribution to the [almost surely unique] random vector that maximizes* Q.

We apply Theorem 6.1 by taking $\xi_i = L_i$, $\theta = d$, $\theta_n = \hat{d}_n$, $\theta_0 = d^0$ and with standardized

$$g(L,d) = (Y - \frac{\beta_\ell^0 + \beta_u^0}{2})(\mathbb{I}_{[X<d]} - \mathbb{I}_{[X<d^0]}).$$

First let's find out the covariance kernel H:

$$\begin{aligned} & Pg(\cdot,\theta_0 + t/\alpha) g(\cdot,\theta_0 + s/\alpha) \\ = \ & \mathbb{E}[(Y - \frac{\beta_\ell^0 + \beta_u^0}{2})^2 (\mathbb{I}_{[X<d+s/\alpha]} - \mathbb{I}_{[X<d]})(\mathbb{I}_{[X<d+t/\alpha]} - \mathbb{I}_{[X<d]})]. \end{aligned}$$

The above expression equals to 0 if s and t are on opposite sides of 0 or $st < 0$. If $st > 0$, it equals

$$\int_{d^0}^{\frac{\min(s,t)}{\alpha} + d^0} p_X(x)dx \int_{-\infty}^{\infty} (y - \frac{\beta_\ell^0 + \beta_u^0}{2})^2 p_\varepsilon(y - f(x))dy.$$

Hence when $st < 0$, $H(s,t) = 0$ and when $st > 0$,

$$\begin{aligned} H(s,t) &= \lim_{\alpha\to\infty} \alpha \int_{d^0}^{\frac{\min(s,t)}{\alpha} + d^0} p_X(x)dx \int_{-\infty}^{\infty} (y - \frac{\beta_\ell^0 + \beta_u^0}{2})^2 p_\varepsilon(y - f(x))dy \\ &= \min(s,t) p_X(d^0) \int_{-\infty}^{\infty} (y - \frac{\beta_\ell^0 + \beta_u^0}{2})^2 p_\varepsilon(y - f(d^0))dy \\ &= \min(s,t) p_X(d^0) \int_{-\infty}^{\infty} (y + f(d^0) - \frac{\beta_\ell^0 + \beta_u^0}{2})^2 p_\varepsilon(y)dy \\ &= \min(s,t) p_X(d^0) \int_{-\infty}^{\infty} y^2 p_\varepsilon(y)dy = \min(s,t) p_X(d^0)\sigma^2, \end{aligned}$$

since $p_X(\cdot)$ is continuous at $x = d^0$ by assumption (i) in (A2).

If the other conditions are satisfied, then, as $n \to \infty$,

$$n^{1/3}(\hat{d}_n - d^0) \to_{\mathcal{D}} W := \text{argmax}_t Q(t),$$

where the limiting process Q is a scaled two-sided Brownian motion, originating from zero, with a quadratic drift:

$$Q(t) = \sigma_0 B(t) - \frac{1}{2} V t^2 = \sigma_0 (B(t) - \frac{1}{2\sigma_0} V t^2),$$

where $\sigma_0^2 = p_X(d^0)\sigma^2$, $B(t)$ is two-sided Brownian motion, and

$$V = -h''(d^0) = -p_X(d^0) f'(d^0) > 0,$$

where $h(d) := Pg(\cdot, d) = \mathbb{E}[(Y - \frac{\beta_\ell^0 + \beta_u^0}{2})\mathbb{I}_{[X<d]}]$; positivity of V is due to the assumption that $h(\cdot)$ has a unique maximizer and the conditions in (A2)(i-ii).

Now let's verify conditions (i-vii) one by one and in order.

Condition (i): Since $P_n g(\cdot, d)$ takes only finite values, this condition is trivially satisfied with an equality.

Condition (ii): The graphs of our function class $\{g(\cdot, d) : d \in (-\infty, \infty)\}$ form a VC class. Hence the class is manageable if it also has a square integrable envelope function. An obvious envelope function is $2|Y - \frac{\beta_\ell^0 + \beta_u^0}{2}|$ and $\mathbb{E}|Y - \frac{\beta_\ell^0 + \beta_u^0}{2}|^2 < \infty$ by assumption (iii) in (A2).

It follows [cf. Pollard, 1990] that almost surely

$$\text{sup}_d |P_n g(\cdot, d) - Pg(\cdot, d)| \to 0.$$

Expanding

$$h(d) = \int_{-\infty}^{\infty} \int_{-\infty}^{d} (y - \frac{\beta_\ell^0 + \beta_u^0}{2}) p_X(x) p_\varepsilon(y - f(x)) dx dy$$

makes it clear that $h(\cdot)$ is continuous by the smoothness conditions in assumption (A2). Because d^0 is the maximizer of $h(\cdot)$,

$$\begin{aligned}
\text{sup}_d |P_n g(\cdot, d) - Pg(L, d)| + h(d^0) &\geq |P_n g(\cdot, \hat{d}_n) - h(\hat{d}_n)| + h(d^0) \\
&\geq |P_n g(\cdot, \hat{d}_n) - h(\hat{d}_n)| + h(\hat{d}_n) \geq P_n g(\cdot, \hat{d}_n) \\
&\geq P_n g(\cdot, d^0) \to h(d^0).
\end{aligned}$$

The last limit holds due to the LLN. It follows that almost surely,

$$\lim_{n\to\infty} h(\hat{d}_n) = h(d^0),$$

which implies that $\hat{d}_n \to d^0$ almost surely, because d^0 is the unique maximizer of $h(\cdot)$ and $h(\cdot)$ is continuous. Hence $\hat{d}_n$ is a consistent estimator of d^0.

Condition (iii): d^0 is an interior point of $\mathcal{D}$ since $\mathcal{D}$ is assumed open and the maximizer is assumed unique.

Now we calculate the envelope function with $\xi = (x, y)$

$$\begin{aligned} G_R(x,y) &:= \sup\{g(x,y,d) : |d - d^0| \le R\} \\ &= \sup_{|d-d^0|\le R}[(y - \frac{\beta_\ell^0 + \beta_u^0}{2})\mathbb{I}_{[x<d]} - \mathbb{I}_{[x<d^0]}] \\ &= |y - \frac{\beta_\ell^0 + \beta_u^0}{2}|\mathbb{I}_{[|x-d^0|<R]}. \end{aligned}$$

$$\begin{aligned} PG_R^2 &= \mathbb{E}(y - \frac{\beta_\ell^0 + \beta_u^0}{2})^2\mathbb{I}_{[|x-d^0|<R]} \\ &= \int_{d^0-R}^{d^0+R}\int_{-\infty}^{\infty} p_X(x)p_\varepsilon(y - f(x))(y - \frac{\beta_\ell^0 + \beta_u^0}{2})^2 dydx \\ &= 2R\, p_X(d^0)\int_{-\infty}^{\infty}(y - \frac{\beta_\ell^0 + \beta_u^0}{2})^2 p_\varepsilon(y - f(d^0))dy(1 + o(1)) \\ &\quad \text{[by the moment conditions in (A2)]} \\ &= O(R). \end{aligned} \tag{6.7}$$

Hence the envelope function is uniformly square integrable for R near 0 and therefore the classes $\mathcal{G}_R$ are uniformly manageable.

Condition (iv): $h(d) := Pg(\cdot, d)$ is twice differentiable at $d = d^0$ because

$$\begin{aligned} h(d) &= \int_{-\infty}^{\infty}\int_{-\infty}^{d} p_X(x)p_\varepsilon(y - f(x))(y - \frac{\beta_\ell^0 + \beta_u^0}{2})dxdy, \\ h'(d) &= \int_{-\infty}^{\infty} p_X(d)p_\varepsilon(y - f(d))(y - \frac{\beta_\ell^0 + \beta_u^0}{2})dy = p_X(d)(f(d) - \frac{\beta_\ell^0 + \beta_u^0}{2}), \\ h''(d) &= p_X'(d)(f(d) - \frac{\beta_\ell^0 + \beta_u^0}{2}) + p_X(d)f'(d). \end{aligned}$$

The existence of the derivatives in the calculation for $h''(d)$ follows from assumptions (i-ii) in (A2). When the maximizer is unique, $f(d^0) = \frac{\beta_\ell^0+\beta_u^0}{2}$. It follows that

$$V = -h''(d^0) = -p_X(d^0)f'(d^0).$$

Condition (v): $H(s,t)$ has been found in the beginning of this proof so that it is enough to verify the second part. For each $\varepsilon > 0$ and $t \in \mathbb{R}^1$,

$$\begin{aligned} &\alpha P[g(\cdot, d^0 + t/\alpha)^2\mathbb{I}_{[g(\cdot,d^0+t/\alpha)>\varepsilon\alpha]} \\ &= \alpha\mathbb{E}(y - \frac{\beta_\ell^0 + \beta_u^0}{2})^2\mathbb{I}_{[x<d^0+t/\alpha]}\mathbb{I}_{[|y-\frac{\beta_\ell^0+\beta_u^0}{2}|>\varepsilon\alpha]} \\ &\le \alpha\mathbb{E}(y - \frac{\beta_\ell^0 + \beta_u^0}{2})^2\mathbb{I}_{[|y-\frac{\beta_\ell^0+\beta_u^0}{2}|>\varepsilon\alpha]} \\ &\le O(\alpha\int_{\varepsilon\alpha}^{\infty} y^2/y^{4+\delta}dy) \ \text{[by the tail condition (iv) in (A2)]} \\ &\le O(\alpha\int_{\varepsilon\alpha}^{\infty} 1/y^{2+\delta}dy) \le O(\alpha/(\varepsilon\alpha)^{1+\delta}dy) \to 0 \quad \text{as } \alpha \to \infty. \end{aligned}$$

Condition (vi): The first part has been shown in (6.7). We now verify the second part. For any $\varepsilon > 0$ and $K > 0$,

$$\begin{aligned} & PG_R^2\{G_R > K\} \\ \leq\ & E(Y - \frac{\beta_\ell^0 + \beta_u^0}{2})^2 \mathbb{I}_{[|X-d^0|<R]} \mathbb{I}_{[|(Y-\frac{\beta_\ell^0+\beta_u^0}{2}|>K]} \\ =\ & \int_{d^0-R}^{d^0+R} p_X(x) \int_{|y-\frac{\beta_\ell^0+\beta_u^0}{2}|>K} |y - \frac{\beta_\ell^0 + \beta_u^0}{2}|^2 p_\varepsilon(y - f(x))dydx \\ \leq\ & M_{p_X} R\ o(1) \ \text{ as } K \to \infty. \end{aligned}$$

The last inequality follows from the fact that both f and p_X are continuous at d^0, hence are bounded by constants M_f and M_{p_X} near d^0 respectively, and from the moment condition (iii) in (A2).

Condition (vii): Without loss of generality, assume $d_1 < d_2$ which are near d^0. Then,

$$|Pg(., d_1) - Pg(., d_2)| \leq M_{p_X} |d_2 - d_1| \int_{-\infty}^{\infty} (|Y| + M_f + |\frac{\beta_\ell^0 + \beta_u^0}{2}|) p_\varepsilon(y)dy,$$

because p_X is bounded near d^0 and the last integral is finite due to the moment condition (iii) in (A2). □

Proof of Fact 3.1:
We provide only a sketch here. It is not hard to show that $\hat{d}_2$ is a consistent estimator of d_2^0 which is the population optimal split point when dividing the original domain of X into two by d_1^0 the limiting point of the first level split. Assume that these two split points d_1^0, d_2^0 are distinct and unique. Because of the consistency of their estimators, without loss of generality, we assume $\hat{d}_2 < \hat{d}_1$. Then,

$$\hat{d}_2 = \text{argmin}_{d_2 < \hat{d}_1} \sum_{i=1}^{n} (Y_i - \beta_{2,\ell}^0 \mathbb{I}_{[X_i < d_2]} - \beta_{2,u}^0 \mathbb{I}_{[X_i \geq d_2]})^2 \mathbb{I}_{[X_i \leq \hat{d}_1]},$$

where $\beta_{2,\ell}^0$ and $\beta_{2,u}^0$ are the best projected values corresponding to the lower partition region. Rewrite

$$\begin{aligned} & (Y_i - \beta_{2,\ell}^0 \mathbb{I}_{[X_i<d_2]} - \beta_{2,u}^0 \mathbb{I}_{[X_i \geq d_2]})^2 - Y_i^2 \\ =\ & (\beta_{2,\ell}^0)^2 \mathbb{I}_{[X_i<d_2]} + (\beta_{2,u}^0)^2 \mathbb{I}_{[X_i \geq d_2]} - 2Y_i \beta_{2,\ell}^0 \mathbb{I}_{[X_i<d_2]} - 2Y_i \beta_{2,u}^0 \mathbb{I}_{[X_i \geq d_2]} \\ =\ & (\beta_{2,\ell}^0 - \beta_{2,u}^0)[(\beta_{2,\ell}^0 + \beta_{2,u}^0) - 2Y_i] \mathbb{I}_{[X_i<d_2]} + \beta_{2,u}^0 (\beta_{2,u}^0 - 2Y_i). \end{aligned}$$

It follows that, assuming $\beta_{2,\ell}^0 > \beta_{2,u}^0$ [without loss of generality]

$$\hat{d}_2 = \text{argmax}_{d_2 < \hat{d}_1} \sum_{i=1}^{n} g(L_i, d_2) \mathbb{I}_{[X_i < \hat{d}_1]}, \quad g(L_i, d_2) = [Y_i - \frac{\beta_{2,\ell}^0 + \beta_{2,u}^0}{2}] \mathbb{I}_{[X_i < d_2]}.$$

Moreover,

$$\sum_{i=1}^{n} g(L_i, d_2) \mathbb{I}_{[X_i < \hat{d}_1]} = \sum_{i=1}^{n} g(L_i, d_2) \mathbb{I}_{[X_i < d_1^0]} + \Delta,$$

where

$$\begin{aligned}\Delta &= \sum_{i=1}^{n} g(L_i, d_2)\mathbb{I}_{[X_i<\hat{d}_1]} - \sum_{i=1}^{n} g(L_i, d_2)\mathbb{I}_{[X_i<d_1^0]} \\ &= \sum_{i=1}^{n}[Y_i - \frac{\beta_{2,\ell}^0 + \beta_{2,u}^0}{2}]\mathbb{I}_{[X_i<d_2]}[\mathbb{I}_{[X_i<\hat{d}_1]} - \mathbb{I}_{[X_i<d_1^0]}].\end{aligned}$$

Because $\hat{d}_1$ converges to d_1^0 and $\hat{d}_2$ converges to d_2^0, and d_1^0 and d_2^0 are distinct, so with high probability,

$$\mathbb{I}_{[X_i<d_2]}(\mathbb{I}_{[X_i<\hat{d}_1]} - \mathbb{I}_{[X_i<d_1^0]}) = 0$$

for d_2 in a neighborhood of d_2^0. That is, $\Delta = 0$ for d_2 in a neighborhood of d_2^0 and with a high probability. It follows that with high probability,

$$\hat{d}_2 = \text{argmax}_{d_2<\hat{d}_1} \sum_{i=1}^{n} g(L_i, d_2)\mathbb{I}_{[X_i<\hat{d}_1]} = \text{argmax}_{d_2<d_1^0} \sum_{i=1}^{n} g(L_i, d_2)\mathbb{I}_{[X_i<d_1^0]}.$$

Comparing with (6.6), we have just shown that $\hat{d}_2$ will have the same asymptotic distribution [but with possibly different distribution parameters] as the estimator for the first level split. The key in this argument is that the unstable regions are non-overlapping when the tree is 'finite' relative to the sample size.

Acknowledgments: We thank Andreas Buja, Keith Knight, an associate editor and a referee for helpful comments. The main work in this paper was conducted when B. Yu was at Bell Labs, Lucent Technologies at Murray Hill (on leave from Berkeley). Moreover, partial supports to B. Yu are acknowledged by Grant DMS-9803063 from NSF and Grants DAAG55-98-1-0341 and DAAD19-01-1-0643 from ARO.

References

[1] Amit, Y. and Geman, D. (1997). Shape quantization and recognition with randomized trees. Neural Computation **9**, 1545–1588.

[2] Bauer, E. and Kohavi, R. (1999). An empirical comparison of voting classification algorithms: bagging, boosting, and variants. Machine Learning **36**, 105–139.

[3] Bickel, P.J., Götze, F. and van Zwet, W.R. (1997). Resampling fewer than n observations: gains, losses, and remedies for losses. Statistica Sinica **7**, 1–32.

[4] Breiman, L. (1996a). Bagging predictors. Machine Learning **24**, 123–140.

[5] Breiman, L. (1996b). Heuristics of instability and stabilization in model selection. Ann. Statist. **24**, 2350–2383.

[6] Breiman, L., Friedman, J.H., Olshen, R.A. & Stone, C.J. (1984). Classification and Regression Trees. Wadsworth, Belmont, CA.

[7] Bühlmann, P. and Yu, B. (2000). Discussion on Additive logistic regression: a statistical view of boosting, auths. J. Friedman, T. Hastie and R. Tibshirani. Ann. Statist. **28**, 377–386.

[8] Buja, A. and Stuetzle, W. (2000a). The effect of bagging on variance, bias, and mean squared error. Preprint. AT&T Labs-Research.

[9] Buja, A. and Stuetzle, W. (2000b). Smoothing effects of bagging. Preprint. AT&T Labs-Research.

[10] Chan K.S. and Tsay, R.S. (1998). Limiting properties of the least squares estimator of a continuous threshold autoregressive model. Biometrika **85**, 413–426.

[11] Dietterich, T.G. (1996). Editorial. Machine Learning **24**, 91–93.

[12] Freedman, D.A. (1981). Bootstrapping regression models. Ann. Statist. **9**, 1218–1228.

[13] Freund, Y. and Schapire, R.E. (1998). Discussion on Arcing classifiers, auth. L. Breiman. Ann. Statist. **26**, 824–832.

[14] Friedman, J.H. (1991). Multivariate adaptive regression splines (with Discussion). Ann. Statist. **19**, 1–67 (Disc: 67–141).

[15] Friedman, J.H. and Hall, P. (2000). On bagging and nonlinear estimation. Preprint.

[16] Giné, E. and Zinn, J. (1990). Bootstrapping general empirical measures. Ann. Probab. **18**, 851–869.

[17] Groeneboom, P. (1989). Brownian motion with a parabolic drift and Airy functions. Probab. Th. Rel. Fields **81**, 79–109.

[18] Hastie, T., Tibshirani, R. and Buja, A. (1994). Flexible discriminant analysis by optimal scoring. J. Amer. Statist. Assoc. **89**, 1255–1270.

[19] Kim, J. and Pollard, D. (1990). Cube root asymptotics. Ann. Statist. **18**, 191–219.

[20] Loh, W.-Y. and Shih, Y-S. (1997). Split selection methods for classification trees. Statistica Sinica **7**, 815–840.

[21] Pollard, D. (1990). Empirical processes : theory and applications. NSF-CBMS regional conference series in probability and statistics, v. 2.

[22] Serfling, R.J. (1980). Approximation Theorems of Mathematical Statistics. Wiley, New York.