Explaining bagging

Explaining Bagging

by

Peter Bühlmann and Bin Yu

Research Report No. 92
May 2000

Seminar für Statistik

Eidgenössische Technische Hochschule (ETH)

CH-8092 Zürich

Switzerland

Explaining Bagging

Peter Bühlmann
Seminar für Statistik
ETH Zentrum
CH-8092 Zürich, Switzerland
and

Bin Yu
Bell Laboratories, Lucent Technologies
and
University of California at Berkeley

May 2000

Abstract

Bagging is one of the most effective computationally intensive procedures to improve on instable estimators or classifiers, useful especially for high dimensional data set problems. Here we formalize the notion of instability and derive theoretical results to explain a variance reduction effect of bagging (or its variant) in 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 contrast with the second-order explanation of Friedman and Hall (2000) for smooth functionals which does not cover the most popular base learner for bagging, namely decision trees.

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 finite many splits.

Heading: Explaining 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 instable estimation/classification schemes. It is extremely useful for large, high dimensional data set problems where finding a good model/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

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

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.

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

(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 'instable'. 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 'instable' 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 'instable 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* $\hat{\theta}_n = h_n(L_1, \ldots, L_n)$ *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 instable predictors as defined above. They arise mainly when hard decisions with indicators are involved as in decision trees [see sections 2 and 3].

The only theoretical investigation on why bagging works is Friedman and Hall (2000). They argue that, for a class of smooth estimators, the first order or leading variance term in an asymptotic analysis remains unchanged under bagging, but the second order variance term is improved (effects on the mean squared error are not studied). This is unsatisfactory since simulation and empirical studies show improvements too large for the second order explanation, for example when sample size is large. Despite of the fact that Friedman and Hall (2000)'s framework is nonlinear, it excludes the prominent case of decision trees: this seriously limits the scope of explanation of bagging provided by their work.

For non-smooth and instable 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 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 and analyze an original predictor that is a tree 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 propose *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 appeared in Friedman and Hall (2000) but without much justification. 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]. 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. We also touch upon classification where the set-up can be somewhat different. In the two-class case and under the assumption that the predictor has favorable 'classification bias', small-order subagging asymptotically drives the misclassification rate to the Bayes rate as subsample size $m \to \infty$ with $m = o(n)$. Such an optimality result with subagging does not hold in the regression case.

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 $\hat{\theta}_n(x)$ allowing an expansion into linear and higher order terms, Friedman and Hall (2000) show that bagging reduces the variance of the only the higher order but *not* of the leading first order asymptotic linear term. Moreover, they do not analyze theoretically the bias term and thus not the mean squared error.

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) does not apply. We particularly consider predictors involving indicator functions. They arise whenever a hard decision is made. For example, 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

$$\hat{\theta}_n(x) = \mathbf{1}_{[\hat{d}_n \le x]}, \ x \in \mathbb{R}, \tag{2.1}$$

where $\hat{d}_n$ is a real-valued estimator based on data $L_i = (Y_i, X_i)$ $(i = 1, \ldots, n)$. If $\hat{d}_n$ is asymptotically normal at rate $b_n^{-1} \to 0$,

$$b_n(\hat{d}_n - d^0) \to_D \mathcal{N}(0, \sigma_\infty^2)$$

with an asymptotic variance $0 < \sigma_\infty^2 < \infty$. Then for an x in the b_n^{-1}-neighborhood of the parameter d^0

$$x = x_n(c) = d^0 + c\sigma_\infty \, b_n^{-1}, \tag{2.2}$$

we have the distributional approximation

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

Denoting by $\Phi(\cdot)$ the c.d.f. of a standard normal distribution, it follows that

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

Thus, $\hat{\theta}_n(x_n(c))$ is inconsistent since the variance does not converge to zero. Moreover, it is instable 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, the bagged predictor

$$\hat{\theta}_{n;B}(x_n(c)) = \mathbb{E}^*[\mathbf{1}_{[b_n(\hat{d}_n^* - \hat{d}_n)/\sigma_\infty \le b_n(x_n(c) - \hat{d}_n)/\sigma_\infty]}] \overset{\mathcal{D}}{\approx} \Phi(c - Z), \; Z \sim \mathcal{N}(0,1), \tag{2.5}$$

where the last approximation follows by assuming that the bootstrap works [see (A1) below] and using (2.3). To formally summarize formulae (2.3) and (2.5), we need the following assumption:

(A1) For some increasing sequence $(b_n)_{n \in \mathbb{N}}$,

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

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

(A1) requires only that the bootstrap works. Due to the results in Giné and Zinn (1990), this essentially holds by assuming i.i.d. observations and $\hat{d}_n$ a smooth functional evaluated at the empirical distribution.

Proposition 2.1 *Assume (A1). For the predictor in (2.1) with $x = x_n(c)$ as in (2.2),*

$$\begin{aligned} &\hat{\theta}_n(x_n(c)) \to_D g(Z) = \mathbf{1}_{[Z \le c]}, \\ &\hat{\theta}_{n;B}(x_n(c)) \to_D g_B(Z) = \Phi(c - Z), \end{aligned}$$

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

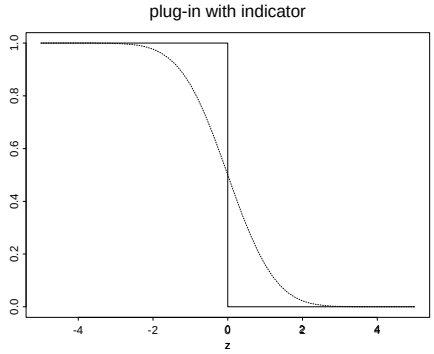

Figure 2.1: Indicator predictor from (2.1) at $x = x_n(0) = d^0$ as in (2.2). Function $g(z) = \mathbf{1}_{[z \le 0]}$ [solid line] and $g_B(z)$ [dotted line] defining the asymptotics of the predictor and its bagged version [see Proposition 2.1].

The distributional approximation of $\hat{\theta}_n(x_n(c))$ is $g(Z)$ with $g(z) = \mathbf{1}_{[z \le c]}$ a *hard-threshold* function; the bagged analogue is $g_B(Z)$ with $g_B(z) = \Phi(c - z)$ a *soft-threshold* function. Figure 2.1 illustrates the two functions $g(\cdot)$ and $g_B(\cdot)$. Bagging reduces variance due to the smoothing or soft- instead of hard-thresholding operation.

Before giving more details, let us consider the instructive case where $x = x_n(0) = d^0$; i.e., x is exactly at the most instable location, where $\mathrm{Var}(\hat{\theta}_n(x))$ is maximal. Proposition 2.1 gives

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

Thus,

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

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 below that for a whole range where $c \neq 0$ in (2.2) [i.e., $x \neq d^0$], bagging still reduces variance while adding only little to the bias.

We compute now the first two asymptotic moments in the instable 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 (2.1) with $x = x_n(c)$ as in (2.2),*

(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. □

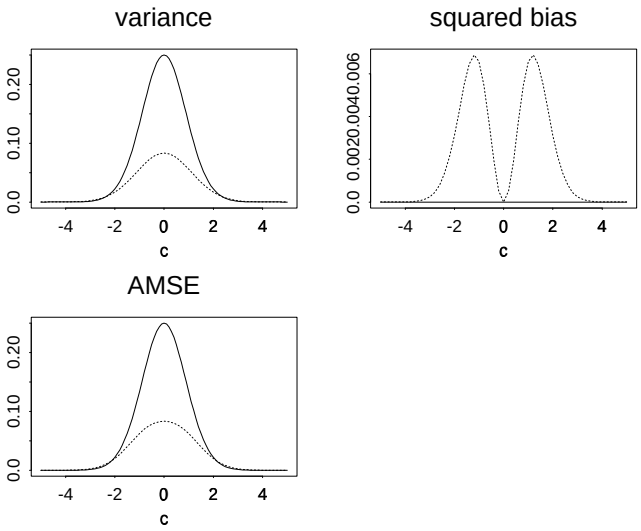

Figure 2.2: Indicator predictor from (2.1) at $x = x_n(c)$ as in (2.2). Asymptotic variance, squared bias and mean squared error [AMSE] for the predictor $\hat{\theta}_n(x_n(c))$ from (2.1) [solid line] and for the bagged predictor $\hat{\theta}_{n;B}(x_n(c))$ [dotted line] as a function of c.

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 instable point $x = d$, 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

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 the 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

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

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 $\hat{\beta}_n$, the MSE is asymptotically additive with p individual MSEs. We thus consider without loss of generality the predictor

$$\hat{\theta}_n(x) = \hat{\beta}\,\mathbf{1}_{[|\hat{\beta}|>u_n]}\,x,\ x \in \mathbb{R}^1, \tag{2.6}$$

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

$$\begin{aligned} &Y_i = \beta X_i + \varepsilon_i,\ X_1, \ldots, X_n\ \mathbb{R}\text{-valued and i.i.d. with } \mathbb{E}|X_i|^2 = 1, \\ &\{\varepsilon_i\}_i \text{ i.i.d. and independent from } \{X_i\}_i,\ \mathbb{E}[\varepsilon_i] = 0,\ \mathrm{Var}(\varepsilon_i) = \sigma^2 < \infty. \end{aligned} \tag{2.7}$$

The threshold is assumed to be of the form

$$u_n = u_n(c) = c\sigma n^{-1/2}. \tag{2.8}$$

Under the null-model $\beta = 0$, this choice leads to a stable predictor $\hat{\theta}_n(x)$ according to Definition 1.2. But instability arises when scaling the predictor $\hat{\theta}_n(x) = \hat{\beta}\,\mathbf{1}_{[|\hat{\beta}|>u_n]}\,x$ with $n^{1/2}$ which becomes an interesting case for bagging.

Proposition 2.2 *Assume model (2.7) with $\beta = 0$ and $\mathbb{E}|\varepsilon_i|^4 < \infty,\ \mathbb{E}|X_i|^4 < \infty$. For the predictor in (2.6) with $u_n = u_n(c)$ as in (2.8),*

$$\begin{aligned} &n^{1/2}\sigma^{-1}\hat{\theta}_n(x) \to_D g(Z) = (Z - Z\,\mathbf{1}_{[|Z|\le c]})x, \\ &n^{1/2}\sigma^{-1}\hat{\theta}_{n;B}(x) \to_D g_B(Z), \end{aligned}$$

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

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

Proof: See 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. 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 $u_n(c)$ [similarly as in Corollary 2.1 and using uniform integrability]: the results are displayed in Figure 2.4. The gain with bagging is quite substantial around $c = \Phi^{-1}(0.975) = 1.96$ which arises for n large under two-sided t-testing on significance level 5%. The bias and mean squared error are here defined for estimating the quantity $\beta x \equiv 0$ since $\beta = 0$; this is different from the centering in section 2.1 where the original predictor is assumed to be asymptotically unbiased [this version can also be deduced from the information in Figure 2.4]. In this particular situation, bagging even has smaller asymptotic bias [for most values of c]; but the bias effect plays again a negligible role in terms of MSE.

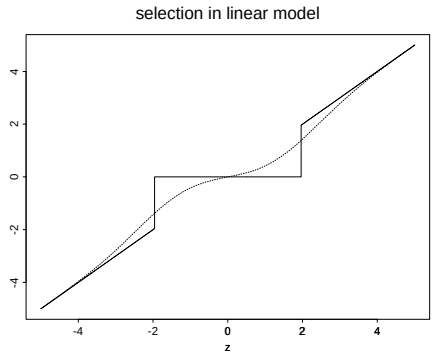

Figure 2.3: Predictor from (2.6) in linear model (2.7). Solid line: function $g(z)$ from Proposition 2.2, defining the asymptotics of $\hat{\theta}_n(x_n(0))$. Dotted line: function $g_B(z)$ from Proposition 2.2, defining the asymptotics of $\hat{\theta}_{n;B}(x_n(0))$.

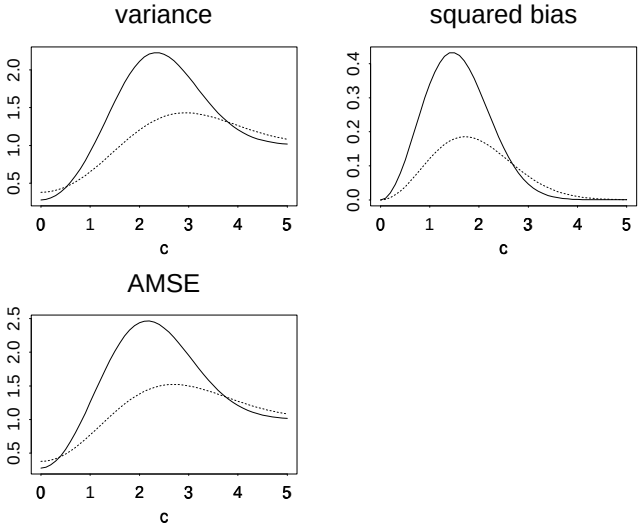

Figure 2.4: Predictor from (2.6) with threshold $u_n = u_n(c)$ from (2.8) in linear model (2.7). Asymptotic variance, squared bias and mean squared error [AMSE], standardized by the factor $n\sigma^{-2}$, as a function of c. Solid line: predictor $\hat{\theta}_n(1)$. Dotted line: bagged predictor $\hat{\theta}_{n;B}(1)$.

2.3 MARS: a soft decision algorithm

We investigate here the effect of bagging on the soft decision algorithm MARS [Friedman, 1991]. For a one-dimensional predictor space, the basic function in MARS is a piecewise linear spline function $[x-d]_+ = (x-d)\,\mathbf{1}_{[d\le x]}$. Its estimated version takes the form

$$\hat{\theta}_n(x) = \hat{\beta}_n[x-\hat{d}_n]_+, \tag{2.9}$$

with the least squares estimates

$$(\hat{\beta}_n, \hat{d}_n) = \text{argmin}_{\beta,d} \sum_{i=1}^{n} (Y_i - \beta[X_i - d]_+)^2 \tag{2.10}$$

for the best projected values

$$(\beta^0, d^0) = \text{argmin}_{\beta,d}\, \mathbb{E}[(Y - \beta[X-d]_+)^2]. \tag{2.11}$$

These estimators behave differently from the hard decision algorithms in a crucial way. We illustrate it in the regression model,

$$Y_i = f(X_i) + \varepsilon_i,\ \text{supp}(X_i) = \mathcal{D} \subseteq \mathbb{R}^1 \text{ an open set},\ \text{supp}(\varepsilon_i) = \mathbb{R}\ \ (i = 1, \ldots, n), \tag{2.12}$$

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

Proposition 2.3 *Consider the regression model (2.12) 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 (2.10) are asymptotically independent and*

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

where β^0, d^0 are as in (2.11).

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

Proposition 2.4 *Under the assumptions of Proposition 2.3, the bootstrap works:*

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

with $\sigma_\beta^2,\ \sigma_d^2$ from Proposition 2.3.

Proof: 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. □

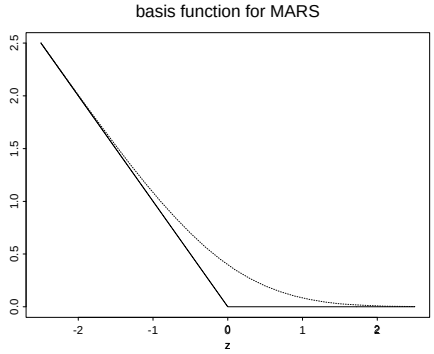

Figure 2.5: MARS basis function. Solid line: function $g(z)$ from Proposition 2.5 defining the asymptotics of $\hat{\theta}_n(x_n(0))$. Dotted line: function $g_B(z)$ defining the asymptotics of $\hat{\theta}_{n;B}(x_n(0))$.

Following Proposition 2.3, the MARS predictor (2.9) 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 (2.6). Due to the hard decision in the latter case, bagging brought in an 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,

$$x = x_n(c) = d^0 + c\sigma_d n^{-1/2}. \tag{2.13}$$

We can write

$$\hat{\theta}_n(x_n(c)) = \beta^0(x_n(c) - \hat{d})_+ + O_P(n^{-1}), \tag{2.14}$$

due to the convergence properties of $\hat{\beta}, \hat{d}$ and the neighborhood definition of $x_n(c)$. The smoothing effect of bagging with MARS is described below.

Proposition 2.5 *Under the conditions of Proposition 2.3,*

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

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

Proof: The first assertion is immediate from (2.14) and Proposition 2.3. The second assertion follows by using Proposition 2.4 and analogously to the proof of Proposition 2.2 in section 6; in particular, we use again formula (6.3). □

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 2.5 [assuming enough moment conditions of the data]. In contrast, for the continuous MARS decision [see Figure 2.5], the bagging improvement is almost negligible.

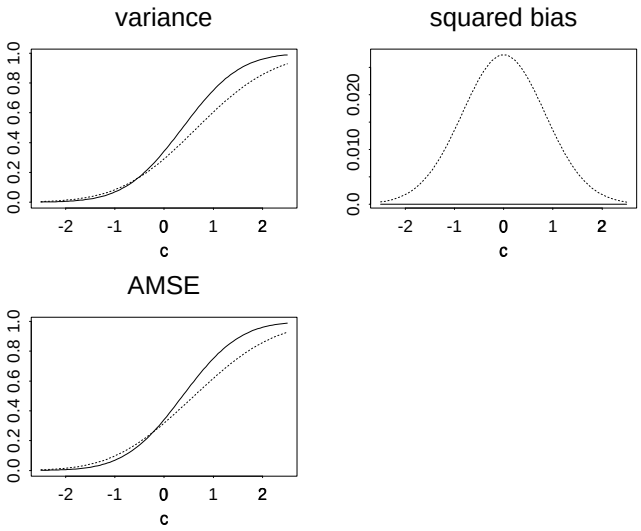

Figure 2.6: MARS predictor $\hat{\theta}_n(x_n(c))$ from (2.9) with $x_n(c)$ from (2.13). Asymptotic variance, squared bias and mean squared error [AMSE], standardized by the factor $n\sigma_d^{-2}$, as a function of c. Solid line: predictor $\hat{\theta}_n(x_n(c))$. Dotted line: bagged predictor $\hat{\theta}_{n;B}(x_n(c))$.

The results for the basic MARS predictor (2.9) 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 for the standard $n^{-1/2}$-rate and non-differentiable estimators: it improves very little in the case of the continuous-decision MARS procedure, but very much upon procedures involving hard, discontinuous decisions.

3 Subagging decision trees

A slight extension of the simple example in section 2.1 is a predictor of the form

$$\hat{\theta}_n(x) = \hat{\beta}_\ell \, \mathbf{1}_{[x<\hat{d}_n]} + \hat{\beta}_u \, \mathbf{1}_{[x\geq\hat{d}_n]} = \hat{\beta}_\ell + (\hat{\beta}_u - \hat{\beta}_\ell) \, \mathbf{1}_{[\hat{d}_n\leq x]},$$

where $\hat{\beta}_\ell$ and $\hat{\beta}_u$ are estimated constants for the regions where x is lower and 'upper' than $\hat{d}_n$, respectively. If these $\hat{\beta}_\ell$ and $\hat{\beta}_u$ are converging in probability to their expectation, they can be asymptotically treated as fixed and the results of section 2.1 apply [for x as in (2.2)], provided that $\hat{d}_n$ is asymptotically normal.

Decision trees such as CART are of the above type, with the predictor

$$\hat{\theta}_n(x) = \sum_{j=1}^{J} \hat{\beta}_j \, \mathbf{1}_{[x\in\mathcal{R}_j]}, \;\; x \in \mathbb{R}^p,$$

where $\{\mathcal{R}_j : j = 1, \ldots, J\}$ is a partition of $\mathbb{R}^p$ and $\hat{\beta}_j = \sum_{i=1}^n Y_i \, \mathbf{1}_{[X_i\in\mathcal{R}_j]} \, / \sum_{i=1}^n \mathbf{1}_{[X_i\in\mathcal{R}_j]}$. The partition cells $\mathcal{R}_j$ are random and the information that x belongs to $\mathcal{R}_j$ can be written as product of indicators of the form

$$\mathbf{1}_{[x\in\mathcal{R}_j]} = \mathbf{1}_{[x^{(i_1)}</\geq\hat{d}_{n,1}]} \, \mathbf{1}_{[x^{(i_2)}</\geq\hat{d}_{n,2}]} \cdots \mathbf{1}_{[x^{(i_{k_j})}</\geq\hat{d}_{n,k_j}]}$$

for some $i_1, \ldots, i_{k_j} \in \{1, \ldots, p\}$, $k_j \in \mathbb{N}$ and some estimators $\hat{d}_{n,i}$ $(i = 1, \ldots, k_j)$; $< / \geq$ denotes either one or the other relation, and $x^{(j)}$ the j-th component of x. However, the asymptotic normality assumption for $\hat{d}_{n,i}$ in (A1) from section 2 fails to be true. For a simple decision tree we give in the next section a rigorous non-standard result, which generalizes to more general decision trees.

3.1 Cube-root asymptotics for the one-split stumps

For a one-dimensional predictor space, a non-normal limit 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 (2.12), consider now the decision tree predictor with stumps,

$$\hat{\theta}_n(x) = \hat{\beta}_\ell \, \mathbf{1}_{[x < \hat{d}_n]} + \hat{\beta}_u \, \mathbf{1}_{[x \geq \hat{d}_n]}, \tag{3.1}$$

where the estimates are obtained by least squares as

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

The best projected values are defined by

$$(\beta_\ell^0, \beta_u^0, d^0) = \operatorname{argmin}_{\beta_\ell, \beta_u, d} \, \mathbb{E}[(Y - \beta_\ell \, \mathbf{1}_{[X < d]} - \beta_u \, \mathbf{1}_{[X \geq d]})^2]. \tag{3.3}$$

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 (2.12). When $\hat{\beta}_\ell$ and $\hat{\beta}_u$ are consistent for β_ℓ^0 and β_u^0, the asymptotics of the predictor in (3.1) for x in a neighborhood of d^0 is equivalent to $\tilde{\theta}_n(x) = \beta_\ell^0 \, \mathbf{1}_{[x \leq \hat{d}]} + \beta_u^0 \, \mathbf{1}_{[x > \hat{d}]}$. To proceed, we make the following assumptions for model (2.12).

(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 and ε 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] $\mathbb{E}[\varepsilon] = \int_{-\infty}^{\infty} y p_\varepsilon(y) dy = 0$, $\sigma^2 = \int_{-\infty}^{\infty} y^2 p_\varepsilon(y) dy < \infty$;

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

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* $n \to \infty$*,*

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

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* $V = -p_X(d^0)f'(d^0) \neq 0$*.*

Proof: see section 6.

Remark 3.1. Theorem 3.1 generalizes to the case where X in (2.12) is p-dimensional with $p > 1$. All what is required is that the theoretically optimal component $\iota^0 \in \{1, \ldots, p\}$ 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 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 empirical LS objective function involves an indicator function being not smooth itself. The $n^{-1/3}$-asymptotics in Theorem 3.1 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 'population' distribution is discrete.

It is worth-noting that (A1) about bootstrap consistency is not necessary for bagging to work as long as the resulted bagged estimator is sensible itself. Conditional on the original sample or the 'population' for the bootstrapped samples, $\hat{d}_n^*$ spreads around d^0 by taking one of the discrete values in the original sample. The resulted bagged stump estimator is a weighted average of the stump estimators with split points at the original sample values of X_i. So $\mathbb{E}^*[\mathbf{1}_{[\hat{d}_n^* \leq \cdot]}]$ is still a smooth thresholding 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 (2.1) 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 $m \leq n$ and $\mathbb{E}|h_m(L_1, \ldots, L_m)(x)|^2 < \infty$ for all x. Then, for any x,*

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

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). □

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. The choice $a = 1/2$ is also considered by Friedman and Hall (2000), mainly in simulations.

We assume now the following very mild condition.

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

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

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

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

We evaluate expectation and variance of subagged estimators for the predictors in (2.1) and (3.1) at instable 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 (2.1) or (3.1) with $x = x_n(c)$ as in (2.2) or (3.6), respectively. Assume that (A3) holds. Suppose $m = [an]$ with $0 < a < 1$. Then,

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

where $\beta_\ell^0 = 0,\ \beta_u^0 = 1$ for the predictor in (2.1).

Proof: See 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.

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, for any $c \in \mathbb{R}$,

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

where $\beta_\ell^0 = 0,\ \beta_u^0 = 1$ for the predictor in (2.1).

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 $b_n = Cn^{1/3}$. □

Numerical evaluation for the 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

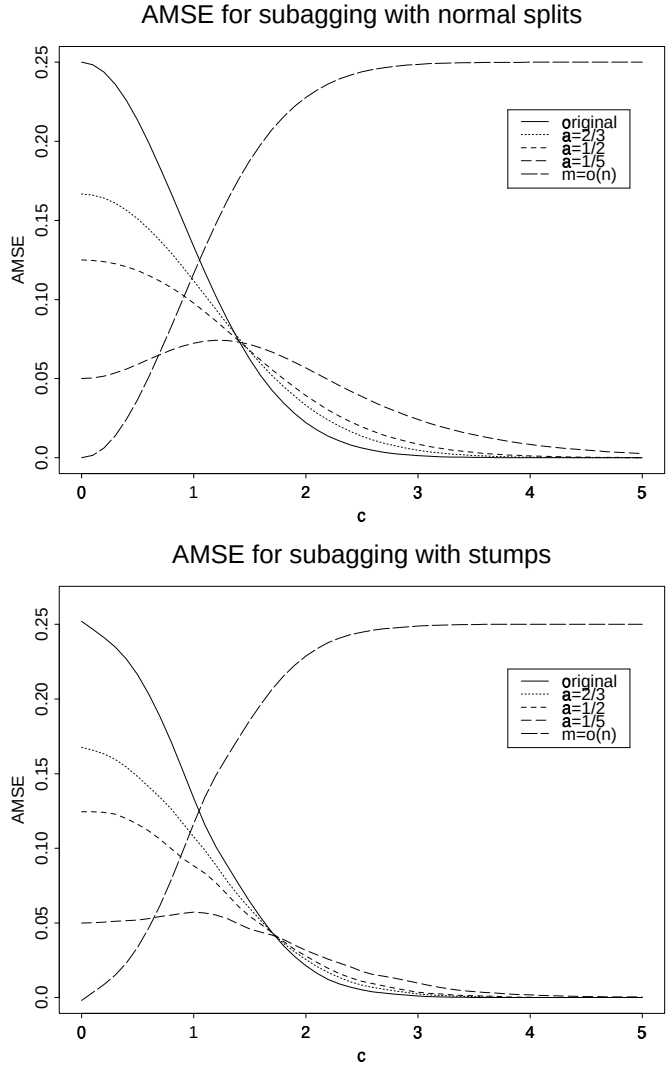

Figure 3.1: Asymptotic mean squared error [AMSE] of original predictor and a bound for the subagged version. Top: indicator predictor $\hat{\theta}_n(x_n(c))$ in (2.1) [solid line] and $\hat{\theta}_{n;SB(m)}(x_n(c))$, with $x_n(c)$ as in (2.2). Bottom: $\hat{\theta}_n(x_n(c))$ in (3.1) [solid line] and $\hat{\theta}_{n;SB(m)}(x_n(c))$, with $x_n(c)$ as in (3.6). In both cases: subsample size $m = [an]$ or $m \to \infty$, $m = o(n)$. The situation corresponds to Theorems 3.2 and 3.3, assuming (A3) with $b_n = Cn^{1/3}$ and $G(\cdot)$ from Theorem 3.1. Everything scaled to $\beta_\ell^0 = 0$, $\beta_u^0 = 1$.

$a = 1/5$ can already become quite bad for 'weak instable 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 [which displays a representative case; the asymptotics depends to a minor degree on various characteristics of the true underlying data-generating mechanism]: fraction subagging with $a = 1/5$ is not behaving poorly at 'weak instable regions' but improves very much at 'strong instable 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 instable 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 a small bandwidth leading to small bias but large variance, and vice versa. From this view, small order subagging has a bias being too large.

3.2.3 Small order subagging in classification

Aggregation in classification is often found empirically to improve similarly as in regression, see also section 4. However, from a mathematical perspective there is at least one noticeable exception as described in the sequel.

Consider the two class problem: the data consists of the pairs $L_i = (Y_i, X_i)$ $(i = 1, \ldots, n)$ but now with binary responses $Y_i \in \{0, 1\}$ and explanatory variables $X_i \in \mathbb{R}^p$. The task is to classify a new variable Y based on its corresponding explanatory X. Given $X = x$, we wish to minimize the following misclassification risk for a classifier $\mathcal{C}(\cdot)$,

$$\mathrm{MCR}(x) = \lambda_0 \mathbb{P}[\mathcal{C}(x) = 1, Y(x) = 0] + \lambda_1 \mathbb{P}[\mathcal{C}(x) = 0, Y(x) = 1],\ \lambda_0, \lambda_1 > 0,$$

where λ_0, λ_1 are the misclassification losses. The classifier is chosen to be of the form [as an estimated version of the optimal Bayes classifier],

$$\hat{\mathcal{C}}_n(x) = \mathbf{1}_{[\hat{P}_n(x) > \lambda]},\ \lambda = \lambda_0/(\lambda_0 + \lambda_1), \tag{3.7}$$

where $\hat{P}_n(x)$ is an estimate of $P(x) = \mathbb{P}[Y = 1|X = x]$. (Su-)bagging of the classifier can be constructed by voting [Breiman, 1996a] or as in another version [cf. Amit and Geman, 1997]

$$\hat{\mathcal{C}}_{n;SB(m)} = \mathbf{1}_{[\hat{P}_{n;SB(m)}(x) > \lambda]},\ \lambda = \lambda_0/(\lambda_0 + \lambda_1),$$

with $\hat{P}_{n;SB(m)}$ as in the regression case and analogously for bagging instead of subagging.

We focus here exclusively on the case where $\hat{P}_n(\cdot)$ is given by stumps with a one-dimensional explanatory variable, as described in section 3.1. The regression technique is applied for an estimate of $\mathbb{E}[Y|X = x] = \mathbb{P}[Y = 1|X = x] = P(x)$ and the estimator thereof is as in (3.1). This is suitable if tree models are used for $\hat{P}_n(\cdot)$, cf. Hastie, Tibshirani and Buja (1994). As in the regression case, we consider the case where $x = x_n(c)$ is in a neighborhood of d^0, described in (3.6). Then, the probability estimator $\hat{P}_n(x_n(c))$ is instable and hence potentially also the classifier $\hat{\mathcal{C}}_n(x_n(c))$. At stable locations with x not in a neighborhood of d, (su-)bagging doesn't change the stump classifier in the first order.

Next a condition is introduced to ensure that the asymptotic stump predictor has classification potential; otherwise, the classifier is asymptotically trivial.

(A4) $\beta_\ell^0 \le \lambda$ and $\beta_u^0 > \lambda$ for the projected values in (3.3), $\lambda = \lambda_0/(\lambda_0 + \lambda_1)$.

The case with $\beta_\ell^0 \ge \lambda$ and $\beta_u^0 < \lambda$ instead of (A4) leads to analogous results, but the notation would need to be given separately. Moreover, we need the following assumption for the classification case as in (A2).

(A5) (i) the conditional distribution $P(x) = \mathbb{P}[Y = 1|X = x]$ exists; $P(\cdot)$ is continuous, and its first and second derivatives P', P'' exist, are uniformly bounded in a neighborhood of d^0 and $P'(d^0) \neq 0$;

(ii) the marginal density p_X of X exists with support being an open set $\mathcal{D} \subseteq \mathbb{R}$; the first derivative p_X' exists, is uniformly bounded in a neighborhood of d^0 and $p_X(d^0) \neq 0$.

Theorem 3.4 *[Small order subagging for classification]*
Consider a classifier $\hat{\mathcal{C}}_n(x)$ as in (3.7) with $\hat{P}_n(x)$ from stumps in (3.1) with $x = x_n(c)$ from (3.6). Assume (A4) and (A5). Suppose the best projected values $\beta_\ell^0, \beta_u^0, d^0$ from (3.3) are unique and subsample size $m \to \infty$, $m = o(n)$. Denote by $MCR_{n;SB(m)}(\cdot)$ and $MCR_n(\cdot)$ the small order subagging stumps and the original stumps misclassification rate, respectively. Let $\lambda = \lambda_0/(\lambda_0 + \lambda_1)$. Then,

$$MCR_{n;SB(m)}(x_n(c)) = \mathbf{1}_{[\beta_\ell^0 + (\beta_u^0 - \beta_\ell^0)G(0) > \lambda]}(\lambda_0 - P(d^0)(\lambda_0 + \lambda_1)) + \lambda_1 P(d^0) + o(1),$$
$$MCR_n(x_n(c)) = G(c)(\lambda_0 - P(d^0)(\lambda_0 + \lambda_1)) + \lambda_1 P(d^0) + o(1).$$

Proof: See section 6.

Note that a rigorous analysis for bagging or fraction subagging seems very difficult at this point. One needs the entire distribution of the asymptotic bootstrap (or fraction subsampling) for $\hat{P}_n^*(\cdot)$ which is highly nontrivial due to non-standard cube-root asymptotics of the original $\hat{P}_n(\cdot)$. For regression, only the first two moments are involved for MSE calculations.

We now give an extended discussion about Theorem 3.4. We always denote in the sequel by $\lambda = \lambda_0/(\lambda_0 + \lambda_1)$. If $P(d^0) = \lambda$, small order subagging does not change the performance. This is actually a consequence of a general fact:

$$\text{any classifiers MCR } (x_n(c)) \text{ converges to the Bayes-MCR}(d^0), \tag{3.8}$$

provided that $P(d^0) = \lambda$ and $P(\cdot)$ continuous. This follows from formula (6.4) in section 6. A situation where this occurs is as follows: $P(\cdot)$ is continuous, point-symmetric around 1/2, i.e. $P(x) = 1 - P(-x)$ and $P(0) = 1/2$, and the design density $p_X(\cdot)$ for X is symmetric around 0, i.e. $p_X(x) = p_X(-x)$. Then, $d^0 = 0$. It is then natural to choose $\lambda_0 = \lambda_1 = 1/2$ as a priori probabilities since the overall chance for an event $Y = 0$ equals 1/2 and assuming equal costs for misclassification. Therefore, $P(d^0) = P(0) = 1/2 = \lambda$. Note that the condition $P(d^0) = \lambda$ describes that classification [even with the true $P(\cdot)$] is not sharp at $x = d^0$: the classifier in (3.7) could as well be defined with the relation '$\ge$', instead of '$>$', which would lead in this case to the opposite result.

The interesting result is thus for $P(d^0) \neq \lambda$. In most reasonable cases, the indicator $\mathbf{1}_{[\beta_\ell^0 + (\beta_u^0 - \beta_\ell^0)G(0) > \lambda]}$ in the MCR for subagging from Theorem 3.4 then 'flips to the advantageous side': that is, $\text{sign}(\lambda - P(d^0)) = \text{sign}\big(\lambda - (\beta_\ell^0 + (\beta_u^0 - \beta_\ell^0)G(0))\big)$.

Theorem 3.5 *Assume the situation in Theorem 3.4 with sign*$(\lambda - P(d^0)) = sign\left(\lambda - \left(\beta_\ell^0 + (\beta_u^0 - \beta_\ell^0)G(0)\right)\right)$*, where* $\lambda = \lambda_0/(\lambda_0 + \lambda_1)$*. Denote by* $MCR_{n;SB(m)}(\cdot)$ *and* $MCR_{Bayes}(\cdot)$ *the small order subagging stumps and Bayes misclassification rate, respectively. Then, for any* $c \in \mathbb{R}$*,*

$$MCR_{n;SB(m)}(x_n(c)) = MCR_{Bayes}(d^0) + o(1).$$

Proof: The Bayes classifier at $x_n(c)$ is $\mathbf{1}_{[P(x_n(c))>\lambda]} \to \mathbf{1}_{[P(d^0)>\lambda]}$ if $P(d^0) \neq \lambda$ [for $P(d^0) = \lambda$, use (3.8)]. The result then follows from (6.4) and Theorem 3.4. □

Small-order subagging can thus be asymptotically optimal among all possible classifiers [provided the assumptions in the above theorem hold]! This result fails to be true with bagging or fraction subagging, since the random fluctuations in $\hat{P}_{n;B}(\cdot)$ or $\hat{P}_{n;SB(m)}(\cdot)$ do not disappear at instable locations. The condition $\text{sign}(\lambda - P(d^0)) = \text{sign}\left(\lambda - (\beta_\ell^0 + (\beta_u^0 - \beta_\ell^0)G(0))\right)$ can be interpreted as 'classification bias' equaling zero: $\beta_\ell^0 + (\beta_u^0 - \beta_\ell^0)G(0)$ is the asymptotic value of $\hat{P}_{n;SB(m)}(x_n(c))$ and we only need it to be on the same side as $P(d^0)$ relative to λ. We know that small order subagging induces an estimation bias term, as described in Theorem 3.3: but this estimation bias might be small enough to cause zero 'classification bias'.

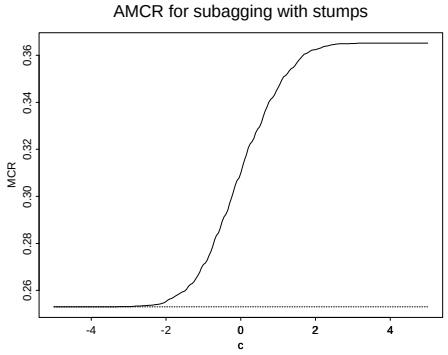

Figure 3.2: Classifier from (3.7) with $\lambda_0 = 0.55, \lambda_1 = 0.45$ and stumps from (3.1) at $x = x_n(c)$ from (3.6). Asymptotic MCR$(x_n(c))$ for original [solid line] and small order subagged [dotted line] classifier as a function of c. The dotted line represents also the Bayes MCR at $x_n(0) = d$. The underlying model is as in (4.1).

We now argue that the condition about 'classification bias' equaling zero, i.e. $\text{sign}(\lambda - P(d^0)) = \text{sign}\left(\lambda - \beta_\ell^0 + (\beta_u^0 - \beta_\ell^0)G(0))\right)$ is plausible. We consider a particular example [the argument easily translates to other examples]. Assume that $P(x) = 1 - P(x)$ with $P(0) = 1/2$. But now, suppose that the design density p_X is symmetric around some point $-s$ $(s > 0)$, i.e. $p_X(s - x) = p_X(s + x)$. Then, $d^0 < 0$. Moreover, assuming equal costs for misclassification, we choose the misclassification losses λ_0, λ_1 as reasonable a-priori probabilities for the events $Y = 0$ and $Y = 1$, respectively. Due to symmetry of $P(\cdot)$ around 0 and of $p_X(\cdot)$ around $-s$, a reasonable choice satisfies $\lambda_0 > 1/2, \lambda_1 = 1 - \lambda_0 < 1/2$. Again due to symmetry, $\beta_\ell^0 < 1 - \beta_u^0$ [or $\beta_\ell^0 + \beta_u^0 < 1$]; since $G(0) = 1/2$, we then obtain $\beta_\ell^0 + G(0)(\beta_u^0 - \beta_\ell^0) = 1/2(\beta_\ell^0 + \beta_u^0) < 1/2 < \lambda_0 = \lambda$; this is the 'correct side' since $P(d^0) < \lambda$. This example is simulated in section 4 to see whether small order subagging approaches the Bayes rate for finite sample sizes; compare also with Figure 3.2.

3.3 Discussion

All the quantifications in the above 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 (2.7) 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

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.

and half subagging are almost identical; a fact which we 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 belong to the partition cell to be refined. Figure 3.3 with

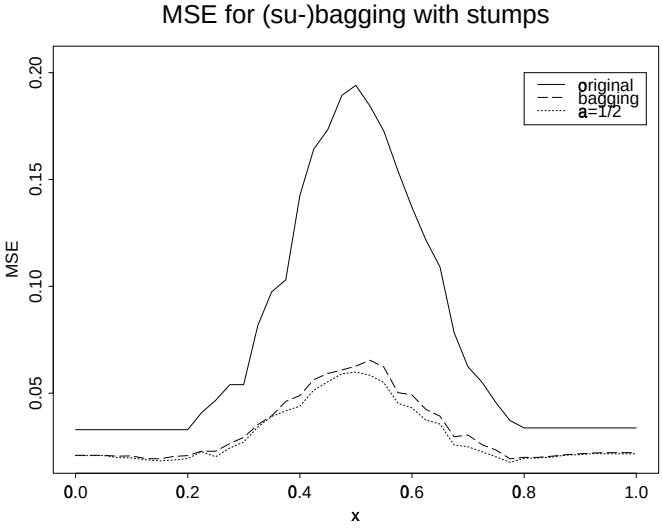

Figure 3.3: Mean squared error of stumps $\hat{\theta}_n(x)$ in (3.1) [solid line] and its (su-)bagged version $\hat{\theta}_{n;SB(m)}(x)$ for $x \in [0, 1]$. Sample size $n = 100$ and subsampling size $m = [an]$. Everything multiplied by the factor $1/(\beta_u^0 - \beta_\ell^0)^2 = 1/2.25$ to obtain [asymptotically] the scale from Figure 3.1.

$n = 100$ is qualitatively as 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. In other words, the bounds can be too conservative. For this case, we get a bound of about 50% on the MSE reduction around the most instable point $c = 0$ while the actual reduction is about 65%.

Our theoretical analysis and its numerical illustrations have only been dealing with a somewhat limited notion of bias. We have always given an a priori advantage to the unbagged predictor and considered performance for estimating $\theta(x) = \lim_{n\to\infty} \mathbb{E}[\hat{\theta}_n(x)]$. (Su-)bagging adds a small bias from this view. We decompose, even for general $\theta(x)$,

$$\mathrm{b}_{n;SB(m)}(x) = \mathbb{E}[\hat{\theta}_{n;SB(m)}(x)] - \theta(x) = (\mathbb{E}[\hat{\theta}_{n;SB(m)}(x)] - \mathbb{E}[\hat{\theta}_n(x)]) + (\mathbb{E}[\hat{\theta}_n(x)] - \theta(x)).$$

The first term reflects the bias due to aggregation with subagging, the second term represents the bias of the original predictor being independent of m. If $m \le n$ increases, the bias $|b_{n;SB(m)}(\cdot)|$ decreases. One way to see this is from

$$\mathrm{b}_{n;SB(m)}(x) = \mathbb{E}[h_m(L_1, \ldots, L_n)(x)] - \mathbb{E}[h_n(L_1, \ldots, L_n)(x)],$$

where $\hat{\theta}_n(x) = h_n(L_1, \ldots, L_n)(x)$: naturally, this is now expected to decrease in absolute value as m increases. Another explanation is given through inspection of Theorem 3.2 and also Theorem 3.3. 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 **o**f **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

$$L_1^*, \ldots, L_m^* \; i.i.d. \sim \hat{F}_n, \tag{3.9}$$

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

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

where $\hat{\theta}_n(x) = h_n(L_1, \ldots, L_n)(x)$. 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 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 $p > 1$.

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 data point. 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. Consider first the case 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, $m_{SB}(v) < m(v)$ 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

$$\mathrm{MSE}_{n;SB}/\mathrm{MSE}_n \sim \int m_{SB}(v) dv / \int m(v) dv, \tag{3.10}$$

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.

A general binary decision tree with k splits is given by a sequence of component indices $\hat{\imath}(i) \in \{1, \ldots, p\}$ and a sequence of splitting variables $\hat{d}_i \in \mathbb{R}$ $(i = 1, \ldots, k)$: the ith split, the component $\hat{\imath}(i)$ of $\mathbb{R}^p$ and the threshold $\hat{d}_i$, then describes the decision to make a refinement of the previous partition, cf. Breiman et al. (1984).

Let us first consider a two split [three terminal node] decision tree in the case of a 1-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 $\mathcal{R}_u = \{x : x \geq \hat{d}_1\}$. The second split $\hat{d}_2$ is defined as

$$(\hat{\beta}_{2;\ell}, \hat{\beta}_{2;u}, \hat{d}_2) = \mathrm{argmin}_{d_2 < \hat{d}_1} \sum_{i=1} (Y_i - (\beta_{2;\ell} \mathbf{1}_{[X_i < d_2]} + \beta_{2;u} \mathbf{1}_{[X_i \geq d_2]} + \hat{\beta}_{1;u} \mathbf{1}_{[X_1 \geq \hat{d}_1]}))^2,$$

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.*

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 [for notational simplicity again with one-dimensional covariates]: 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

$$\int m_{i;SB}(v)dv / \int m_i(v)dv << 1, \ i = 1, 2. \tag{3.11}$$

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 (3.11). The relative gain with (su-)bagging should thus be similar to the one for stumps in (3.10).

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 and thus to the case where decision trees are most popular, see also Remark 3.1.

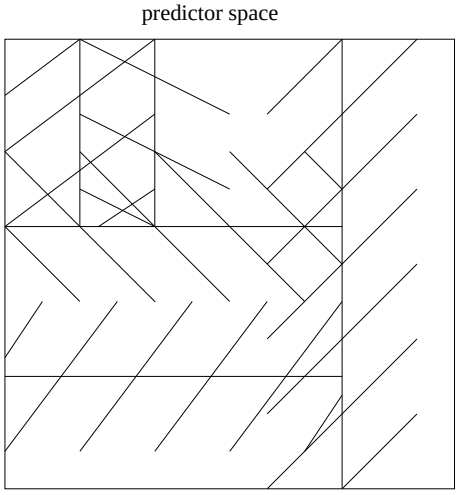

Figure 3.4: Schematic representation of instability regions in 2-dimensional predictor space. 5 binary splits [indicated by horizontal or vertical lines], instability regions [indicated by diagonals].

We now argue that a large decision tree is strongly instable. Suggested by Theorem 3.1 and Fact 3.1, every splitting variable $\hat{d}_i$ has a range $A_{n,i}$ of instability in the predictor space,

$$A_{n,i} = \{x \in \mathbb{R}^p : x^{(\hat{\imath}(i))} = d_i^0 + c\sigma_{i,\infty} n_i^{-1/3},\ c \in \mathbb{R}\}. \tag{3.12}$$

Thereby, d_i^0 denotes the best projected value of the i-th splitting variable, $\sigma_{i,\infty}^2$ the limiting variance of $\hat{d}_i$, and n_i the number of observations involved for determining this split. We have implicitly assumed here that the asymptotic quantities are well defined. Already a moderate number of instability regions $A_{n,1}, A_{n,2}, \ldots, A_{n,k}$ fill out the predictor space $\mathbb{R}^p$, see Figure 3.4 [note that also with $p > 2$ dimensions, instability regions are large; they are p-dimensional subsets with one coordinate diameter $O(n^{-1/3})$ and all others infinite]. The reason for this is given by the relatively large diameter for a coordinate of instability region $A_{n,i}$ behaving as $O(n_i^{-1/3})$ with $n_i \leq n$. [We recognize that we are using our asymptotic results to situations where n_i have orders such as 10. The legitimacy of such a use is supported, to a certain extent, by the simulation results in Table 3.1 for $n = 10$, in a special model]. Figure 3.2 exploits this empirically: a relatively large instability region for $n = 100$ with stumps having one-dimensional predictor space $[0, 1]$ [with uniform design for the explanatory variables]. We then conclude that 'very many points' in the predictor space are instability point. Instability of hard threshold decision trees has been exploited from a different view also by Loh and Shih (1997).

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

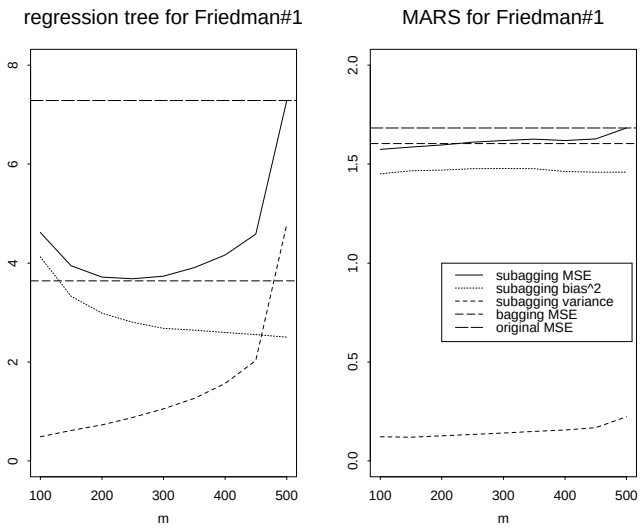

Figure 4.1: Performance for a large regression tree and MARS and their (su-)bagged versions in the simulated model Friedman #1.

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]) distributions. 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 by 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 [optimal MSE 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 with 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 them over 50 training-test-set random divisions.

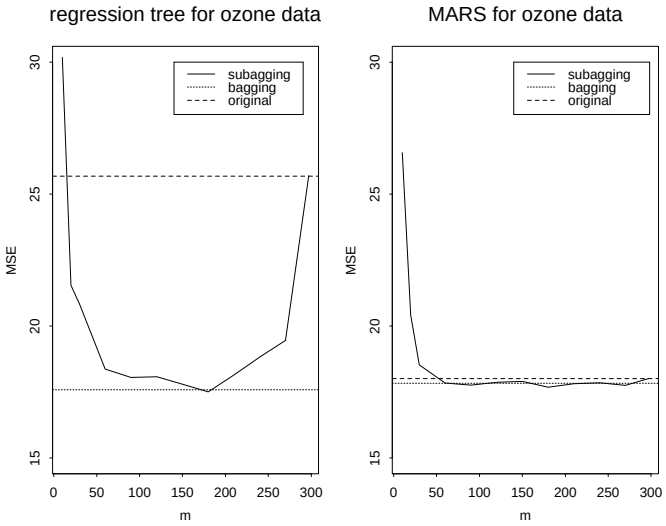

Figure 4.2: Mean squared error performance for a large regression tree and MARS and their (su-)bagged versions for the ozone data.

Figure 4.2 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 first a simulated example which corresponds to the situation described at the end of subsection 3.2.2, and to Figure 3.2,

$$\begin{aligned} &Y_1, \ldots, Y_n \text{ independent, } Y_i \sim \text{Bernoulli}(P(X_i)), \ \log(P(x)/(1-P(x))) = x, \\ &X_1, \ldots, X_n \text{ i.i.d. } \sim \mathcal{N}(-0.25, 1). \end{aligned} \tag{4.1}$$

Sample sizes are $n = 100$ and $n = 500$. The numbers of simulations over the model and replicates for aggregation are 100 and 50, respectively. Figure 4.3 displays the results. As expected from Theorem 3.5 [the model (4.1) implies the conditions of Theorem 3.5], subagging with a small subsample size m comes close to the Bayes MCR and outperforms bagging, and more clearly the original classifier.

We consider here also 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 is estimated with random division in training-test-sets, analogously as for the MSE with the ozone data set in the previous section. [The misclassification rate is $\mathbb{P}[\mathcal{C}(X) \neq Y(X)]$, i.e. equal misclassification losses]. Figure 4.4 displays the results. Bagging is slightly better than half subagging. This is one of the examples showing among the worst [but still small] magnitude of loss with subagging compared to bagging: relatively large subsample sizes are needed for good performance [maybe due

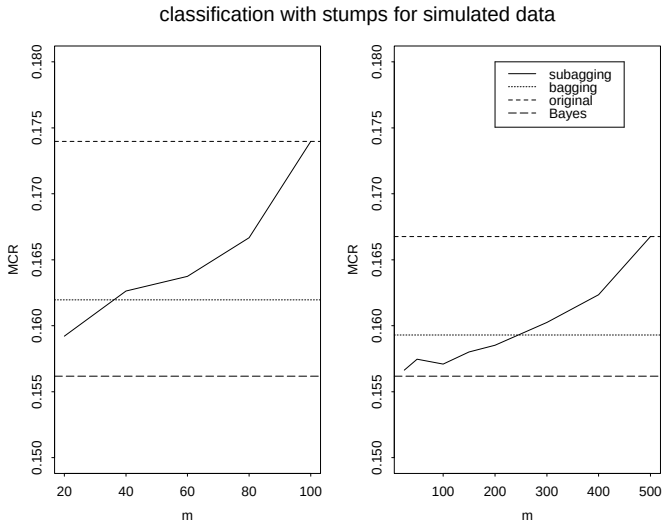

Figure 4.3: Misclassification rate [MCR] for the classifier in (3.7) with $\hat{P}_n(\cdot)$ a stump and its (su-)bagged version. The data is described by (4.1). Left panel: sample size $n = 100$. Right panel: sample size $n = 500$.

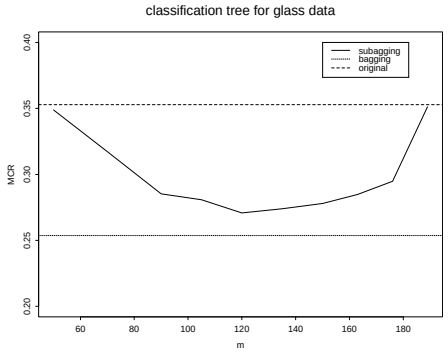

Figure 4.4: Misclassification rate [MCR] for the classifier in (3.7) with $\hat{P}_n(\cdot)$ a large tree and its (su-)bagged version.

to the relatively small training sample size, namely 189, for trees with many splits], in contrast to Figure 4.3.

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 instable in the sense of Definition 1.2. Generally, (su-)bagging doesn't make the predictor stable, but it stabilizes to a certain extent: with one mathematically noticeable exception where small order subagging for classification stabilizes completely to the optimal Bayes predictor [provided that assumptions (A4) and (A5) holds]! 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. 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 how (su-)bagging works for 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 recognized before. For example, Breiman (1996a) and Dietterich (1996) [in his second implication] exclusively mention high dimensional schemes. 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]. The computational advantage of subagging can be even compounded with a better performance in classification where small order subagging can become optimal. 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 $\mathbb{E}^*|\hat{\beta}^*|^2 = O_P(1)$],

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

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

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

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

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

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

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 $\mathrm{Var}(h_m(L_1, \ldots, L_m)(x_n(c)))$. □

Proof of Theorem 3.4:
Under assumption (A5), arguments similar to the proof of Theorem 3.1 below establish the $n^{1/3}$-asymptotics of $\hat{d}_n$. The misclassification rate MCR of a classifier $\mathcal{C}$ can be rewritten as,

$$\mathrm{MCR}(x) = \pi(x)(\lambda_0 - P(x)(\lambda_0 + \lambda_1)) + \lambda_1 P(x), \tag{6.4}$$

where $\pi(x) = \mathbb{P}[\mathcal{C}(x) = 1]$. For the original classifier,

$$\mathbb{P}[\hat{\mathcal{C}}_n(x_n(c)) = 1] = \mathbb{P}[\hat{d}_n \le x_n(c)] + o(1) = G(c) + o(1),$$

according to (A4), Theorem 3.1 and the definition of $x_n(c)$ in (3.6). For the subagged classifier, $\mathbb{P}[\hat{\mathcal{C}}_{n;SB(m)}(x_n(c)) = 1] = \mathbb{P}[\hat{P}_{n;SB(m)}(x_n(c)) > \lambda]$ ($\lambda = \lambda_0/(\lambda_0 + \lambda_1)$) follows from Theorem 3.3. This completes the proof. □

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

$$\hat{d}_n = \text{argmin}_d \sum_{i=1}^{n} (Y_i - \beta_\ell^0 \, \mathbf{1}_{[X_i<d]} - \beta_u^0 \, \mathbf{1}_{[X_i \geq d]})^2.$$

Rewrite

$$\begin{aligned} & (Y_i - \beta_\ell^0 \, \mathbf{1}_{[X_i<d]} - \beta_u^0 \, \mathbf{1}_{[X_i \geq d]})^2 - Y_i^2 \\ = \; & (\beta_\ell^0)^2 \, \mathbf{1}_{[X_i<d]} + (\beta_u^0)^2 \, \mathbf{1}_{[X_i \geq d]} - 2Y_i \beta_\ell^0 \, \mathbf{1}_{[X_i<d]} - 2Y_i \beta_u^0 \, \mathbf{1}_{[X_i \geq d]} \\ = \; & (\beta_\ell^0 - \beta_u^0)[(\beta_\ell^0 + \beta_u^0) - 2Y_i] \, \mathbf{1}_{[X_i<d]} + \beta_u^0 (\beta_u^0 - 2Y_i). \end{aligned}$$

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

$$\hat{d}_n = \text{argmax}_d \sum_{i=1}^{n} g(L_i, d), \quad g(L_i, d) = (Y_i - \frac{\beta_\ell^0 + \beta_u^0}{2}) \, \mathbf{1}_{[X_i<d]} \, . \tag{6.5}$$

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

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

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) $H(s,t) = \lim_{\alpha\to\infty} \alpha Pg(\cdot,\theta_0 + t/\alpha)g(\cdot,\theta_0 + s/\alpha)$ *exists for each* s,t *in* $\mathbb{R}^k$ *and* $\lim_{\alpha\to\infty} \alpha Pg(\cdot,\theta_0 + t/\alpha)^2\{|g(\cdot,\theta_0 + t/\alpha| > \varepsilon\alpha\} = 0$ *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* $PG_R^2 \mathbf{1}_{[G_R>K]} <> \varepsilon R$ *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

$$g(L,d) = (Y - \frac{\beta_\ell^0 + \beta_u^0}{2})(\mathbf{1}_{[X<d]} - \mathbf{1}_{[X<d^0]}).$$

First let's find out the covariance kernel H:

$$\begin{aligned} & Pg(\cdot,\theta_0 + t/\alpha)g(\cdot,\theta_0 + s/\alpha) \\ = \; & \mathbb{E}[(Y - \frac{\beta_\ell^0 + \beta_u^0}{2})^2(\mathbf{1}_{[X<d+s/\alpha]} - \mathbf{1}_{[X<d]})(\mathbf{1}_{[X<d+t/\alpha]} - \mathbf{1}_{[X<d]})]. \end{aligned}$$

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:

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

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 $h(d) := Pg(\cdot, d) = \mathbb{E}[(Y - \frac{\beta_\ell^0+\beta_u^0}{2})\,\mathbf{1}_{[X<d]}]$; 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

$$\sup_d |P_n g(\cdot, d) - Pg(\cdot, d)| \to 0.$$

Expanding

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

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

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

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 $\xi = (x, y)$

$$\begin{aligned}
G_R(x, y) &:= \sup\{g(x, y, d) : |d - d^0| \leq R\} \\
&= \sup_{|d-d^0|\leq R}[(y - \frac{\beta_\ell^0 + \beta_u^0}{2})\,\mathbf{1}_{[x<d]} - \mathbf{1}_{[x<d^0]}] \\
&= |y - \frac{\beta_\ell^0 + \beta_u^0}{2}|\,\mathbf{1}_{[|x-d^0|<R]}\,.
\end{aligned}$$

$$
\begin{aligned}
PG_R^2 &= \mathbb{E}(y - \frac{\beta_\ell^0 + \beta_u^0}{2})^2 \mathbf{1}_{[|x-d^0|<R]} \\
&= \int_{d^0-R}^{d^0+R} \int_{-\infty}^{\infty} p_X(x) p_\varepsilon(y - f(x))(y - \frac{\beta_\ell^0 + \beta_u^0}{2})^2 dy dx \\
&= 2R\, p_X(d^0) \int_{-\infty}^{\infty} (y - \frac{\beta_\ell^0 + \beta_u^0}{2})^2 p_\varepsilon(y - f(d^0)) dy (1 + o(1)) \\
&\quad \text{[by the moment conditions in (A2)]} \\
&= O(R). \qquad (6.6)
\end{aligned}
$$

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

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

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 $t \in \mathbb{R}^1$,

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

Condition (vi): The first part has been shown in (6.6). We now verify the second part. For any $\varepsilon > 0$ and $K > 0$,

$$
PG_R^2 \{G_R > K\}
$$

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

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,

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

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

$$
\begin{aligned}
& (Y_i - \beta_{2,\ell}^0 \, \mathbf{1}_{[X_i < d_2]} - \beta_{2,u}^0 \, \mathbf{1}_{[X_i \geq d_2]})^2 - Y_i^2 \\
= \; & (\beta_{2,\ell}^0)^2 \, \mathbf{1}_{[X_i < d_2]} + (\beta_{2,u}^0)^2 \, \mathbf{1}_{[X_i \geq d_2]} - 2 Y_i \beta_{2,\ell}^0 \, \mathbf{1}_{[X_i < d_2]} - 2 Y_i \beta_{2,u}^0 \, \mathbf{1}_{[X_i \geq d_2]} \\
= \; & (\beta_{2,\ell}^0 - \beta_{2,u}^0)[(\beta_{2,\ell}^0 + \beta_{2,u}^0) - 2 Y_i] \, \mathbf{1}_{[X_i < d_2]} + \beta_{2,u}^0 (\beta_{2,u}^0 - 2 Y_i).
\end{aligned}
$$

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

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

Moreover,

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

where

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

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,

$$\mathbf{1}_{[X_i<d_2]}(\mathbf{1}_{[X_i<\hat{d}_1]} - \mathbf{1}_{[X_i<d_1^0]}) = 0$$

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,

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

Comparing with (6.5), 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 instable regions are non-overlapping when the tree is 'finite' relative to the sample size.

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. To appear in Ann. Statist.

[8] 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.

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

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

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

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

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

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

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

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

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

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

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

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