

On Bagging and Nonlinear Estimation

Jerome H. Friedman* Peter Hall†

January 4, 2000

Abstract

We study the decomposition of statistical estimators into linear and higher order parts, or equivalently, the decomposition of the objective function that they optimize into quadratic and higher order terms. We show that bagging reduces the variability of the nonlinear component by replacing it with an estimate of its expected value, while leaving the linear part unaffected. It is therefore most successful when used with highly nonlinear estimators such as decision trees and neural networks. We investigate different resampling schemes and show that half-sampling without replacement is virtually equivalent to traditional bootstrap sampling. It is shown that sampling fractions other than 1/2 often work better with bagging, and that there can be a bias-variance trade-off in choosing an optimal value. In addition to reducing variance, bagging is seen to also reduce bias with certain types of estimators that include decision trees.

1 Introduction

Bagging was introduced by Breiman (1996) as a means for improving the accuracy of estimators of functions $\theta(\mathbf{x})$ of a multivariate argument $\mathbf{x} = \{x_1, \dots, x_p\}$ from data $\{y_i, \mathbf{x}_i\}_1^n$

$$\hat{\theta}(\mathbf{x}) = \arg \max_{\theta(\mathbf{x}) \in \Theta} L(\theta(\mathbf{x})). \quad (1.1)$$

Here Θ represents a function class representable by the estimator, such as neural networks or decision trees. The objective function $L(\theta(\mathbf{x}))$ is a data based estimate of the expected value of some functional such as log-likelihood or other negative loss function $l(y, \theta)$,

$$L(\theta(\mathbf{x})) = \frac{1}{n} \sum_{i=1}^n l(y_i, \theta(\mathbf{x}_i)). \quad (1.2)$$

“Bagging” the estimator $\hat{\theta}$ involves repeating (1.1) and (1.2) many times, B , each time on a different randomly drawn subsample $S_b \subset \{y_i, \mathbf{x}_i\}_1^n$ of the data. This induces a series of estimates

$$\hat{\theta}_b(\mathbf{x}) = \arg \max_{\theta(\mathbf{x}) \in \Theta} \frac{1}{n} \sum_{i \in S_b} l(y_i, \theta(\mathbf{x}_i)), \quad b \in \{1, \dots, B\}.$$

The resulting “bagged” estimate is taken to be their average,

$$\hat{\theta}_B(\mathbf{x}) = \frac{1}{B} \sum_{b=1}^B \hat{\theta}_b(\mathbf{x}).$$

*Statistics Department, Stanford University, Stanford, CA 94305; jhf@stat.stanford.edu

†CSIRO Mathematical Sciences and Centre for Mathematics and its Application, Australian National University, Canberra ACT 0200, Australia; peter.hall@anu.edu.au

Since its introduction, considerable evidence has been accumulated that clearly demonstrates the effectiveness of bagging some estimators, such as decision trees and neural networks; it is now routinely used. The underlying reasons for its success have been less clear.

A statistic such as $\hat{\theta}$ (see (1.1) and (1.2)) may often be interpreted as involving a combination of terms of differing degrees of complexity, for example a linear part combined with quadratic, cubic and higher-order components. We argue that the operation of bagging an estimator leaves the linear part unchanged, but reduces variability of the other contributions by replacing them by empirical approximations to their expected values. This ‘algebraic’ view of bagging has a simple geometric analogue: the high-order terms represent stochastic ‘bumps’ on the basic paraboloid shape of the objective function $L(\theta)$, at (1.2), that is optimized by the statistic, and bagging replaces the bumps by an empirical approximation to their average value. That reduces variability of the bumps, with the result that the objective function is better approximated by the paraboloid, whose optimum occurs at the value of the linear component of the statistic.

These views imply that the more linear is an estimator, or the more parabolic is its associated objective function near the optimum, the less effective bagging will be. And *vice versa*, the more effective bagging proves to be, the less linear is the problem. For example, estimators derived from linear least-squares regression, ridge regression, and non-adaptive kernel and nearest-neighbor methods should not receive much variance reduction through bagging. On the other hand, highly nonlinear methods such as decision trees and neural networks should benefit substantially. Some variance reduction methods, such as regularization, have an impact on both linear and nonlinear parts of an estimator. Our interpretations of bagging suggest that regularization can’t be effectively replaced by bagging if the linear component is significant, but that in highly nonlinear problems there may not be much point in using both bagging and regularization.

In terms of relative variability, bagging an estimator moves it closer to its linear approximation, which is generally an unbiased estimate of the population parameter

$$\theta^*(\mathbf{x}) = \arg \max_{\theta(\mathbf{x}) \in \Theta} E_{y, \mathbf{x}} l(y, \theta(\mathbf{x})),$$

with relatively low variance. The latter property arises not just from the fact that a linear function is typically less variable than a higher-degree polynomial, but because the linear component is the one to which optimality theory (e.g. Cramér-Rao efficiency bounds) for an estimation method usually applies. Our geometric interpretation of bagging implies that it also operates to reduce the difficulty that numerical procedures have finding the global optimum of an objective function — by replacing multiple local optima by their average value, their number and heights are reduced, and the surface of the objective function becomes more regular, thereby further reducing variance.

In many problems the algebraic interpretation represents something of an abstraction. The settings where bagging produces greatest improvement in performance are those where mathematical complexity is so great that traditional expansions in linear, quadratic and higher-order terms are valid only formally, and that is the context where we discuss our interpretations. We show that relatively abstract mathematical models, where complexity increases with sample size, can be used to extend the range of a theoretical treatment. And we note that the algebraic interpretation leads to modified bagged estimators that allow quadratic or higher-order contributions to be eliminated entirely.

We summarise our theoretical and geometric ideas in Section 2. In particular, Section 2.5 works through an example which illustrates not only the effect that bagging has on linear and quadratic terms, but also the way in which the impact increases with complexity of the problem. The effects of different types of bootstrapping (e.g. different values of m in the m -out-of- n bootstrap, and with-replacement versus without-replacement algorithms) are studied theoretically in relatively simple settings, and explored by simulation in more complex cases; see Section 3. Conclusions from simple cases are seen to be generally valid in the more complex settings. In particular, ‘without replacement’ bagging, using half samples, produces results that are close in important respects to those obtained using standard ‘with replacement’ bootstrapping. This reflects the fact that they produce virtually identical variance-reduced forms of the quadratic

component of an estimator; the quadratic part often makes the most significant contribution to variability, after the linear one (which is left untouched by both forms of bagging).

Ideas related to our own have been expressed by Breiman (1996, 1999) in his invocation of a ‘fairy godmother’, who gives us an endless supply of samples over which we might average so as to reduce the variability of an estimator. We would qualify that view by arguing that the fairy godmother is only interested in averaging nonlinear parts of the estimator, or the bumps on the paraboloid; she won’t allow us to average linear terms or the paraboloid itself. And we can’t quite achieve the performance of her estimators at even those levels, since we have to make do with averaging operations that are based on the original sample.

With-replacement resampling is of course in the spirit of the contemporary bootstrap. With-out-replacement methods were employed in early approaches to resampling, for example those of Mahalanobis (1946) and Hartigan (1969, 1971). McCarthy (1966) was an early proponent of ‘without replacement’ resampling using half samples. Efron (1982) discussed related issues, including (Efron, 1982, pp. 62–4) the estimation of variance by half sampling. Our mathematical arguments are related to those in a number of other places, for example Efron and Stein’s (1981) account of high-order properties of the jackknife estimate of variance.

2 Interpretations of Bagging

2.1 Taylor-expansion models for estimators.

Many estimators $\hat{\theta}$, whether constructed by bootstrap methods or otherwise, are linear to first order under conventional asymptotic models. However, for small to moderate sample sizes, and particularly in complex or high-dimensional settings, the quadratic and higher-order terms can influence performance. In such cases we may express a vector-valued parameter $\hat{\theta}$ as

$$\hat{\theta} = \bar{X} + \sum_{r_1} \sum_{r_2} (\bar{Y} - \mu)^{(r_1)} (\bar{Y} - \mu)^{(r_2)} a_{r_1 r_2} + \sum_{r_1} \sum_{r_2} \sum_{r_3} (\bar{Y} - \mu)^{(r_1)} (\bar{Y} - \mu)^{(r_2)} (\bar{Y} - \mu)^{(r_3)} a_{r_1 r_2 r_3} + \dots, \quad (2.1)$$

where $\bar{X} = n^{-1} \sum_i X_i$, $\bar{Y} = n^{-1} \sum_i Y_i$, X_i and Y_i are each vector-valued functions of the i ’th datum U_i , bracketed superscripts indicate vector components, $\mu = E(Y_i)$, and the vectors a_{r_1, \dots, r_k} , each having the same length as X_i and $\hat{\theta}$, are nonrandom. We intend (2.1) to be interpreted in an abstract sense, with higher-order terms representing contributions of greater complexity to the estimator $\hat{\theta}$. In those cases where (2.1) is strictly valid, the linear component \bar{X} would generally be unbiased for the true value θ_0 of the parameter θ of which $\hat{\theta}$ is an estimator; see the discussion below. In such settings the linear estimator \bar{X} would represent the asymptotic form of $\hat{\theta}$.

To appreciate the origins of an expansion such as (2.1), suppose $\theta = \hat{\theta}$ is derived by solving an ‘estimating equation’, such as

$$\sum_{i=1}^n g(U_i, \theta) = 0, \quad (2.2)$$

where g is a smooth multivariate function with the same number of components as θ , θ_0 is the solution of $E\{g(U, \theta)\} = 0$, $\mathcal{U} = \{U_1, \dots, U_n\}$ represents the sample, and U denotes a generic datum U_i . We may re-express (2.2), through Taylor expansion of $g(U_i, \theta)$ about θ_0 , as

$$\sum_{i=1}^n \left\{ g(U_i, \theta_0) + \sum_r g_r(U_i, \theta_0) (\hat{\theta} - \theta_0)^{(r)} + \frac{1}{2} \sum_{r_1} \sum_{r_2} g_{r_1 r_2}(U_i, \theta_0) (\hat{\theta} - \theta_0)^{(r_1)} (\hat{\theta} - \theta_0)^{(r_2)} + \dots \right\} = 0, \quad (2.3)$$

where g_{r_1, \dots, r_k} denotes the partial derivative of g with respect to $\theta^{(r_1)}, \dots, \theta^{(r_k)}$. After solving (2.3) for $\hat{\theta}$ we obtain (2.1), with $\bar{X} = \theta_0 + M^{-1} n^{-1} \sum_i g(U_i, \theta_0)$, where M is the square matrix whose r 'th column equals $E\{g_r(U, \theta_0)\}$, and \bar{Y} in (2.1) denotes a vector whose elements are averages, over the index i , of respective components of the quantities $g_r(U_i, \theta_0), g_{r_1 r_2}(U_i, \theta_0), \dots$.

The expansions at (2.1) and (2.3) are generally only asymptotic, meaning that the sizes of the remainders represented by ' \dots ' are of the order of the first omitted terms. In particular, the expansions do not necessarily converge as infinite series of polynomials in $\bar{Y} - \mu$ and $\hat{\theta} - \theta_0$, respectively, and the length of the vector \bar{Y} in (2.1) would depend on the number of terms to which we took that expansion.

2.2 Geometric interpretations of models.

Equation (2.2) usually arises as the vector of derivatives, or scores, of an objective function such as a log-likelihood or a measure of risk or loss, with respect to components of θ . (In nonparametric problems, where $\hat{\theta}$ would be a bootstrap estimator on account of each datum being given equal weight in (2.2), the log-likelihood would be defined under a model of convenience, usually representing only in broad structural terms the true distribution of the data.) The objective function is, to first order, parabolic in shape, with its vertex at \bar{X} . To second order the paraboloid has cubic-polynomial 'bumps' superimposed on it, representing terms of degree 3 in a Taylor expansion of the objective function, and influencing the estimator $\hat{\theta}$ primarily through the quadratic terms in (2.1). A third-order approximation to the shape of the objective function incorporates quartic-degree bumps. The cubic terms in (2.1) are the result of first-order influences of the quartic bumps and second-order influences of the cubic bumps in the objective function.

The effects of these perturbations become more pronounced as the complexity (e.g. dimensionality, or the magnitude of high-order derivatives of the objective function,) of the problem increases, to such an extent that the linear approximation $\hat{\theta} \approx \bar{X}$ may no longer adequately describe properties of the estimator. In such cases, the point $\hat{\theta}$ at which the objective function is optimized can be some distance from the linear component, \bar{X} . The positions of bumps on the paraboloid are of course stochastic, and the heights of the bumps can have large variance since they represent polynomials of relatively high degree.

Of course, this effect would be greatly reduced if the random bumps could be replaced by their average value, or — almost equivalently — if the influence of the bumps could be replaced by its expected value. In algebraic terms, the latter change amounts to altering $\hat{\theta}$, at (2.1), to the value it would take if the quadratic and higher-order terms on the right-hand side were replaced by their expected values:

$$\bar{\theta} = \bar{X} + E \left\{ \sum_{r_1} \sum_{r_2} (\bar{Y} - \mu)^{(r_1)} (\bar{Y} - \mu)^{(r_2)} a_{r_1 r_2} + \sum_{r_1} \sum_{r_2} \sum_{r_3} (\bar{Y} - \mu)^{(r_1)} (\bar{Y} - \mu)^{(r_2)} (\bar{Y} - \mu)^{(r_3)} a_{r_1 r_2 r_3} + \dots \right\}. \quad (2.4)$$

We shall argue in Section 2.3 that 'bagging' the estimator $\hat{\theta}$ produces an empirical approximation to $\bar{\theta}$. The estimator $\bar{\theta}$ is the one that Breiman's (1996, 1999) 'fairy godmother' would use.

Note particularly that bagging leaves unchanged the linear term in the formula for the estimator. That is, it doesn't affect the basic paraboloid shape of the objective function. It averages the

bumps on the paraboloid with one another, or, virtually equivalently, it averages out quadratic and higher-order terms in a Taylor expansion of $\hat{\theta}$. Of course, since bagging reproduces the expectation in (2.4) only in empirical terms, the variance is not actually reduced to that of the linear component of \bar{X} , but it is reduced nevertheless. The more ‘linear’ a statistic is, the less useful bagging will be in improving its performance.

2.3 Basic properties of the bagged estimator.

Different approaches to bagging amount to taking averages, using different sample re-use methods, of values of the estimator $\hat{\theta}$. One approach is the m -out-of- n ‘with replacement’ bootstrap form, $\hat{\theta}_{\text{bag},1} = E(\hat{\theta}^*|\mathcal{U})$, where $\hat{\theta}^*$ denotes the version of $\hat{\theta}$ computed not from \mathcal{U} but from a resample \mathcal{U}^* of size $m \leq n$, drawn by sampling with replacement from \mathcal{U} . Another is the ‘without replacement’ bootstrap form, $\hat{\theta}_{\text{bag},2} = E(\hat{\theta}^\dagger|\mathcal{U})$, where $\hat{\theta}^\dagger$ denotes the version of $\hat{\theta}$ computed from a resample \mathcal{U}^\dagger of size $m \leq n-1$ drawn by sampling without replacement from \mathcal{U} . (We shall also use the ‘asterisk’ and ‘dagger’ notation for other statistics, such as \bar{X} and \bar{Y} .) If $\hat{\theta}$ is linear in functions of the data then $\hat{\theta}_{\text{bag},1} = \hat{\theta}_{\text{bag},2} = \hat{\theta}$, and more generally, $E(\bar{X}^*|\mathcal{U}) = E(\bar{X}^\dagger|\mathcal{U}) = \bar{X}$. Therefore, in the class of problems addressed at (2.1), bagging does not alter the linear component of $\hat{\theta}$; it affects only quadratic and higher-order terms.

Define $\hat{\sigma}_{r_1 r_2} = n^{-1} \sum_i (Y_i - \bar{Y})^{(r_1)} (Y_i - \bar{Y})^{(r_2)}$, $S = \sum_{r_1} \sum_{r_2} \hat{\sigma}_{r_1 r_2} a_{r_1 r_2}$ and $\alpha_m = n/m \geq 1$, and suppose $\hat{\theta}$ admits the representation at (2.1). We claim that if $\alpha_m \rightarrow \alpha$ as $n \rightarrow \infty$, where $1 \leq \alpha < \infty$ and $\alpha > 1$ in the case of $\hat{\theta}_{\text{bag},2}$, then

$$\hat{\theta}_{\text{bag},1} = \bar{X} + n^{-1} \alpha_m S + \delta_{\text{bag},1}, \quad \hat{\theta}_{\text{bag},2} = \bar{X} + n^{-1} (\alpha_m - 1) S + \delta_{\text{bag},2}, \quad (2.5)$$

where the ‘remainder’ terms denoted by $\delta_{\text{bag},1}$ and $\delta_{\text{bag},2}$ represent higher-order perturbations. (An outline derivation of (2.5) is given in Section 2.6.) The terms in S on the right-hand sides in (2.5) derive from the quadratic term on the right-hand side of (2.1). Cubic, etc, terms in (2.1) go into the remainders $\delta_{\text{bag},1}$ and $\delta_{\text{bag},2}$ in (2.5), and have explicit formulae analogous to those for the quadratic terms.

It is clear that the variance of $\hat{\sigma}_{r_1 r_2}$, and hence of S , is of order n^{-1} , and so the variances of the terms immediately after the mean \bar{X} on the right-hand sides in (2.5) are both $O(n^{-3})$. By way of comparison, we may deduce from (2.1) that $\hat{\theta} = \bar{X} + \Delta$ where the variance of Δ is asymptotic to a constant multiple of n^{-2} . Therefore, bagging has reduced the variability of quadratic contributions to the estimator $\hat{\theta}$. It has the same effect on higher-order contributions, such as cubic terms.

At (2.4) we gave an idealized form, $\bar{\theta}$, of the bagged estimator, in which we took the expectation of all terms other than the first, linear one. It is clear from (2.1) and (2.4) that $\bar{\theta}$ admits the expansion

$$\bar{\theta} = \bar{X} + n^{-1} s + \delta, \quad (2.6)$$

where δ represents higher-order perturbations, $s = \sum_{r_1} \sum_{r_2} \sigma_{r_1 r_2} a_{r_1 r_2}$, and $\sigma_{r_1 r_2} = E\{(Y - \mu)^{(r_1)} (Y - \mu)^{(r_2)}\}$, with Y denoting a generic value of Y_i . The first expansion at (2.5) in the case $m = n$, i.e. in the conventional n -out-of- n ‘with replacement’ bootstrap form of bagging, is clearly the standard bootstrap version of (2.6). When $m < n$, in either the with-replacement or without-replacement cases, the factors α_m and $(\alpha_m - 1)$ in (2.5) arise through different averaging operations being used in those cases.

Note that S is virtually linear in the data; in particular, S is linear in the quantities $\hat{\sigma}_{r_1 r_2}$, which are standard bootstrap variance estimators and so are linear to first order. Similar remarks apply to higher-order terms in expansions of $\hat{\theta}_{\text{bag},1}$ and $\hat{\theta}_{\text{bag},2}$. It follows that standard bias correction methods, for example based on the bootstrap or the jackknife, will not significantly alleviate the bias of bagged estimators.

2.4 Taking $m < n$.

Although (2.5) suggests that taking α very close to 1 in the ‘without replacement’ version of bagging might remove entirely the effects of the quadratic term in the estimator $\hat{\theta}$, that interpretation ignores inflationary effects that such a choice has on the remainder, $\delta_{\text{bag},2}$. Our assumptions prior to (2.5) require that in the ‘without replacement’ case, $\alpha_m \rightarrow \alpha > 1$, and so m cannot be nearer to n than order n .

Nevertheless, the second formula at (2.5) implies that for $m \sim (1 - \epsilon)n$, where $0 < \epsilon < \frac{1}{2}$ is fixed and n is large, ‘without replacement’ bagging will produce an estimator with a smaller quadratic term and hence, at least from an asymptotic viewpoint, lesser asymptotic variance. Taking $m \sim \frac{1}{2}n$ in ‘without replacement’ bagging produces an estimator where the effect of the quadratic term is virtually identical to that in n -out-of- n ‘with replacement’ bagging.

There is reason to expect the n -out-of- n ‘with replacement’ bootstrap, and $\frac{1}{2}n$ -out-of- n ‘without replacement’ bootstrap, to perform similarly in a range of settings, not just the present one. Note that the effective size of an n -out-of- n ‘with replacement’ bootstrap resample U^* , in terms of the amount of information it contains, is given by the ratio of the

$$\frac{(\sum_{i=1}^n N_i)^2}{\sum_{i=1}^n N_i^2} \sim \frac{1}{2}n, \quad (2.7)$$

where N_i denotes the number of times the i 'th data value U_i is repeated in U^* . In particular, the variance of the mean of N_i copies of independent and identically distributed random variables Z_i , for $1 \leq i \leq n$, is very nearly equal to twice the variance of the mean of the n independent random variables themselves, for large n .

2.5 Application of theory to specific problems

The Taylor expansion model introduced at (2.1) is of course a simplification, or abstraction, of a range of complex settings to which bagging may be applied. In particular, in many instances the right-hand side of (2.1) would include an additional, linear term, $\sum_r (\bar{Y} - \mu)^{(r)} a_r$ say, which we have dropped in order to simplify our exposition. It should also be pointed out that a reduction in the variance of the the second term on the right-hand side of (2.1) does not necessarily lead to a reduction in the variance of $\hat{\theta}$, although it will if the quantities $a_{r_1 r_2}$ there are adequately large, or, more commonly in complex problems, if there are sufficiently many such terms.

The latter caveat is important: only in relatively complex problems, where the number of explicitly or implicitly fitted parameters (or an equivalent index of complexity) is large, will the quadratic terms play a significant role and so lead to variance reduction. In such cases, (2.1) provides insight into a range of specific, practical problems where bagging offers reduced variability. We shall treat one of them in moderate detail, and discuss others briefly. The case we treat in detail is that of an over-parametrised model, where estimator variance is increased by fitting a relatively large number of parameters. This example is representative of problems, arising for example in highly nonlinear regression, where ‘nuisance parameters’ accommodate effects that are difficult to identify and which can impact more on the variability than on the consistency of estimators of parameters θ that are really of interest. Theoretical arguments based on (2.1) enable us to identify the way in which bagging can reduce variability in this setting, in cases where complexity (in terms of the number of nuisance parameters) is high relative to sample size.

Let $f(x|\omega)$ denote a probability density, supported on the positive half-line and determined by a parameter vector ω which will play the role of the ‘nuisance parameters’. Suppose that for each choice of ω the corresponding distribution has unit mean:

$$\int f(x|\omega) dx = \int x f(x|\omega) dx = 1 \quad \text{for all } \omega.$$

We shall incorporate an extra parameter $\theta > 0$, describing location: $f(x|\theta, \omega) = \theta^{-1} f(x/\theta|\omega)$. Thus, ω determines only the shape of the distribution with density $f(x|\theta, \omega)$, not its location.

For each ω the sample mean \bar{X} is a consistent estimator of θ , but is not necessarily the most efficient. We shall take $\omega = 0$ to represent the case where $f(x|\omega)$ is the standard exponential density, in which case \bar{X} is the maximum likelihood estimator of θ .

More generally, let $(\hat{\theta}, \hat{\omega})$ denote the maximum likelihood estimator of (θ, ω) , and define $\hat{f}(x|\theta) = f(x|\theta, \hat{\omega})$. Then, $\hat{\theta}$ is obtained by maximising $\sum_i \log \hat{f}(X_i|\theta)$ with respect to θ , from which it may be deduced by Taylor expansion that

$$\hat{\theta} - \theta = \sum_{k \geq 1} \psi_k(\hat{\omega}) \prod_{j=1}^k \left\{ n^{-1} \sum_{i=1}^n \chi_{kj}(X_i|\hat{\omega}) \right\},$$

for functions ψ and χ with $\psi_1 \equiv 1$ and each $E\chi(X|\omega) = 0$. (Here and below, ω denotes the true value of that vector.) Now Taylor-expand $\hat{\omega}$ around ω , to prove that for functions ξ ,

$$\begin{aligned} \hat{\theta} - \theta = n^{-1} \sum_{i=1}^n \xi_1(X_i|\omega) + \sum_{r_1} \sum_{r_2} \left\{ n^{-1} \sum_{i=1}^n \xi_{2,r_1}(X_i|\omega) \right\} \\ \times \left\{ n^{-1} \sum_{i=1}^n \xi_{2,r_2}(X_i|\omega) \right\} a_{r_1 r_2} + \dots, \end{aligned} \quad (2.8)$$

say, where each $E\xi(X|\omega) = 0$. Formula (2.8) is a close analogue of (2.1); in the next paragraph we shall show that the formulae can be made even closer.

Suppose that the true value of ω is 0, so that fitting the parameters ω actually degrades performance relative to simply using the sample mean. In this case, $\xi_1(x|\theta) = x - \theta + \eta(x|\theta)$, where $\eta(\cdot|\theta)$ integrates to 0 against the exponential density with mean θ , and vanishes if ω is taken to be empty. Then, (2.8) becomes

$$\hat{\theta} = \bar{X} + n^{-1} \sum_{i=1}^n \eta(X_i|\theta) + \sum_{r_1} \sum_{r_2} \bar{Y}^{(r_1)}(\omega) \bar{Y}^{(r_2)}(\omega) a_{r_1 r_2} + \dots, \quad (2.9)$$

where we have written $\bar{Y}(\omega)$ for the average, over i , of a quantity such as $\xi_{2,r_1}(X_i|\omega)$ at (2.8), and ω would be replaced by the true value 0. Compare (2.9) with (2.1). The number of components in each of the double series at (2.8) and (2.9) equals $(\nu + 1)^2$, where ν is the number of components of ω and the '+1' derives from θ itself.

As we argued in Section 2.3, a major effect of bagging is to reduce the variability of the quadratic term (that is, the third term) in (2.9). If we use conventional asymptotics, where ν is held fixed as $n \rightarrow \infty$, then this will be negligible relative to the variability of the leading terms at (2.9). However, if ν is large then the total variability of the quadratic term can be considerable; note that the number of components of the quadratic term grows virtually in proportion to ν^2 , as ν increases. Therefore, the arguments given earlier in this Section, focusing on quadratic terms, indicate that the capacity of bagging to reduce variance in such problems becomes greater as the number of fitted 'nuisance parameters' — that is, the complexity of the fitted model — increases.

Other examples include estimators that are explicit functions of means, yet can have particularly high variability. A case in point is that where $\hat{\theta} = p(\bar{U})/q(\bar{U})$, where p and q are smooth ℓ -variate functions such as polynomials, and \bar{U} is an ℓ -variate sample mean. Taylor expanding $p(\bar{U})$ and $q(\bar{U})$ about $\mu = E(U)$ we obtain a version of (2.1).

The high variability of $\hat{\theta}$ in this setting usually arises because $q(\bar{U})$ attains values close to 0 with moderately high probability. By working instead with $\hat{\theta}_{\text{bag}} = E(\hat{\theta}^*|\mathcal{U})$ or $E(\hat{\theta}^\dagger|\mathcal{U})$ we often obtain an estimator with lesser variance than $\hat{\theta}$. In more extreme cases, for example where the support of the distribution of $q(U)$ includes the origin, using the conditional median rather than the conditional mean can further reduce variance. Numerical examples in the settings of econometrics and boundary estimation are given by Hall and Simar (1999).

The special cases $\hat{\theta} = \bar{U}^2$ or $(1 + \bar{U}^2)^{-1}$, where \bar{U} is now a univariate mean, serve to illustrate some of the features of this class of problems. It may be shown that in the first of these examples, $\text{var } \hat{\theta}_{\text{bag}} - \text{var } \hat{\theta} = 4\mu\beta n^{-2} + O(n^{-3})$, where $\beta = E(U - \mu)^3$ denotes skewness. In particular, if neither the mean nor the skewness vanishes then, asymptotically, bagging reduces variance if and only if these quantities have opposite signs.

The fact that variance is not reduced more generally is a consequence of the univariate, low-complexity nature of this problem. In the analogue of (2.1) for the square of a mean there is only one quadratic term:

$$\bar{U}^2 = \mu^2 + 2\mu(\bar{U} - \mu) + (\bar{U} - \mu)^2,$$

where successive terms on the right-hand side represent constant, linear and quadratic contributions respectively. Therefore, the opportunity of the quadratic term to contribute to variance is relatively low. Consequently, reducing the variance of the quadratic term has relatively little effect on overall variability. On this occasion it is the ‘interaction’ of the quadratic term with the linear term that dictates the final variance of the bagged estimator.

Similar results hold in the case $\hat{\theta} = (1 + \bar{U}^2)^{-1}$. Here, depending on the sampled distribution, bagging can reduce variance, but not in all circumstances and only to second order. However, asymptotic analyses such as these are very conventional. They fail to address cases where either the problem grows in complexity with increasing sample size, or the sampled distribution has high variance. For example, if the true mean of the sampled distribution is either 0 or sufficiently close to 0, and if the true variance is of order n , then it may be proved that when $\hat{\theta} = (1 + \bar{U}^2)^{-1}$ the bagged estimator generally has lesser variance.

2.6 Derivation of (2.5)

Suppose the quantities a_{r_1, \dots, r_k} appearing at (2.1) may be expressed in the form $b_{r_1, \dots, r_k}(\nu)$, where ν equals the expected value of a function ζ of U , and the function b_{r_1, \dots, r_k} is known. For example, this is the case if $\hat{\theta}$ is derived by inverting the expansion at (2.3). Suppose too that $\hat{\theta}$ is a scalar; in effect we are doing a component-wise analysis of $\hat{\theta}$. Using either ‘with replacement’ or ‘without replacement’ bagging, in the bootstrap world μ should be replaced by the conditional mean of the distribution from which each resampled datum is drawn, which equals \bar{Y} . Likewise, ν should be replaced by $\bar{Z} = n^{-1} \sum_i \zeta(U_i)$, and (2.1) becomes

$$\begin{aligned} \hat{\theta}^* &= \bar{X}^* + \sum_{r_1} \sum_{r_2} (\bar{Y}^* - \bar{Y})^{(r_1)} (\bar{Y}^* - \bar{Y})^{(r_2)} b_{r_1 r_2}(\bar{Z}) + \dots, \\ \hat{\theta}^\dagger &= \bar{X}^\dagger + \sum_{r_1} \sum_{r_2} (\bar{Y}^\dagger - \bar{Y})^{(r_1)} (\bar{Y}^\dagger - \bar{Y})^{(r_2)} b_{r_1 r_2}(\bar{Z}) + \dots \end{aligned}$$

in the respective cases.

Define $\rho_m = (m - 1)/(n - 1)$. Provided $2 \leq m \leq n$ in the ‘with replacement’ case, and $2 \leq m \leq n - 1$ in the ‘without replacement’ case,

$$\begin{aligned} E\{(\bar{Y}^* - \bar{Y})^{(r_1)} (\bar{Y}^* - \bar{Y})^{(r_2)} | \mathcal{U}\} &= m^{-1} \hat{\sigma}_{r_1 r_2}, \\ E\{(\bar{Y}^\dagger - \bar{Y})^{(r_1)} (\bar{Y}^\dagger - \bar{Y})^{(r_2)} | \mathcal{U}\} & \\ &= m^{-2} \left[m(m - 1) E\{(Y_1^\dagger - \bar{Y})^{(r_1)} (Y_2^\dagger - \bar{Y})^{(r_2)} | \mathcal{U}\} \right. \\ &\quad \left. + m E\{(Y_1^\dagger - \bar{Y})^{(r_1)} (Y_1^\dagger - \bar{Y})^{(r_2)} | \mathcal{U}\} \right] \\ &= m^{-2} \left\{ \frac{m(m - 1)}{n(n - 1)} \sum_{i_1 \neq i_2} (Y_{i_1} - \bar{Y})^{(r_1)} (Y_{i_2} - \bar{Y})^{(r_2)} \right. \\ &\quad \left. + \frac{m}{n} \sum_{i=1}^n (Y_i - \bar{Y})^{(r_1)} (Y_i - \bar{Y})^{(r_2)} \right\} \\ &= m^{-1} (1 - \rho_m) \hat{\sigma}_{r_1 r_2}. \end{aligned}$$

These formulae give the second terms in the respective expansions at (2.5). In the case of the second expansion the exact result has $\alpha_m - 1$ at (2.5) replaced by $\alpha_m - (1 - m^{-1})(1 - n^{-1})^{-1}$, but since $n \sim \alpha m$ then the difference goes into the remainder.

The calculations above are given in the case of conditional means of the quadratic terms in expansions of $E(\hat{\theta}^*|\mathcal{U})$ and $E(\hat{\theta}^\dagger|\mathcal{U})$, but virtually identical calculations apply to means of higher-degree terms. In particular, the conditional mean of the polynomial of degree k in an expansion of $\hat{\theta}^*$ or $\hat{\theta}^\dagger$ equals a function of order $n^{-\langle(k+1)/2\rangle}$, with variance of order $n^{-2\langle(k+1)/2\rangle-1}$, where $\langle x \rangle$ denotes the largest integer not exceeding x . For $k \geq 3$ these terms are of lower order than the quadratic contribution.

In conclusion we note that the covariance between the first and second terms in either of the expansions at (2.5) is of order n^{-2} , which is the same as the order of the covariance between the first and second terms on the right-hand side of (2.1). While this is of larger order than the variance of the second term in any of these expansions, in practical applications to complex problems the contribution of the covariance will actually be smaller than that of the variance of the second term. This may be seen, for example, from the discussion in Section 2.5 of the effect of the number of quadratic components on overall variability.

3 Numerical Experiments

In order to gain further insights we present the results of several simulation experiments. All involve estimating a function of a multivariate argument in noisy settings using regression trees. Regression trees (Breiman *et al* 1984) represent a very nonlinear estimation method. Data $\{y_i, \mathbf{x}_i\}_1^n$ were generated according to the model

$$y_i = f(\mathbf{x}_i) + \sigma \varepsilon_i, \quad (3.1)$$

with each \mathbf{x}_i independently generated from a 10-dimensional uniform distribution, $\mathbf{x}_i \sim U^{10}[0, 1]$. Each ε_i was randomly drawn from a standard normal distribution. Three “target” functions $f(\mathbf{x})$ were considered:

- constant: $f(\mathbf{x}) = 0$, $\sigma = 1$,
- piecewise-constant: $f(\mathbf{x}) = \prod_{j=1}^5 1(x_j \geq 0.13)$, $\sigma = 0.5$,
- linear: $f(\mathbf{x}) = \sum_{j=1}^5 j \cdot x_j$, $\sigma = 3$.

Note that the last two targets are functions of only five of the ten predictor variables, so that the others represent irrelevant “noise” variables. Each generated data set was used to induce a 50-terminal node regression tree, producing a corresponding function estimate $\hat{f}(\mathbf{x})$. Average bias-squared

$$B^2 = E_{\mathbf{x}}\{f(\mathbf{x}) - E_{y\mathbf{x}}\hat{f}(\mathbf{x})\}^2 \quad (3.2)$$

and average variance

$$V = E_{y\mathbf{x}}\{\hat{f}(\mathbf{x}) - E_{y\mathbf{x}}\hat{f}(\mathbf{x})\}^2 \quad (3.3)$$

were computed by averaging over 100 independent trials for each experiment. In all experiments $B = 50$ bagging iterations were employed.

3.1 Constant target

In this setting we study the effect of various forms of bagging on variance alone, since here regression trees are unbiased. Figure 1 shows the average variance (3.3) of the bagged regression tree estimator $\hat{f}(\mathbf{x})$ as a function of the sampling fraction m/n , without (left frame) and with (right

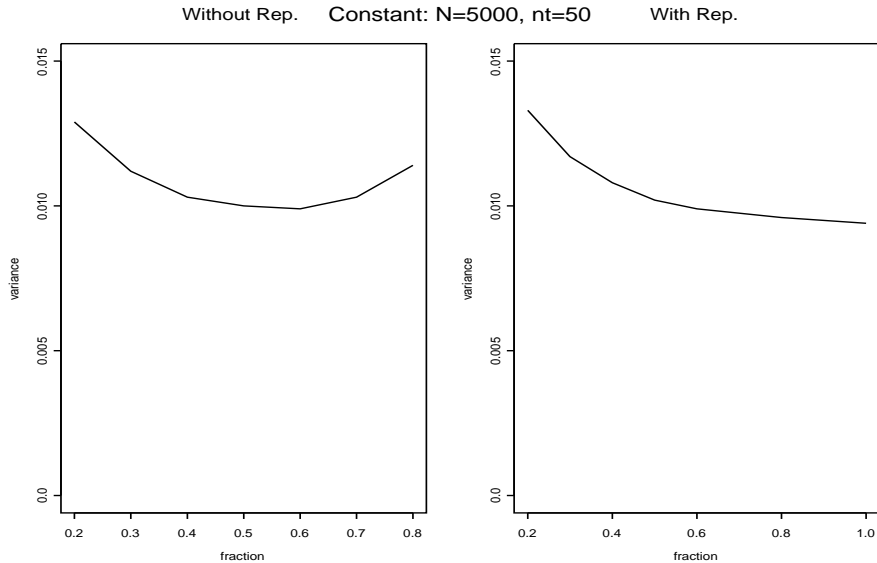


Figure 1: Average variance as a function of sampling fraction m/n without (left frame) and with (right frame) replacement, for $n = 5000$, with a constant target. The variance of the unbagged estimate is 0.0856.

frame) replacement. Training samples of $n = 5000$ were used to induce the trees. The variance of the original “unbagged” estimate was 0.0856. Thus, for all sampling fractions shown, both types of bagging dramatically reduce variance. The optimal sampling fraction is approximately 0.6 for without replacement sampling and 1.0 for sampling with replacement. However, the curves are fairly flat near their optima, so that a choice is not critical. Note that smaller fractions require less computation.

These results reflect the theoretical conclusions reached in Section 2, in that (a) the variance of without-replacement bagging is approximately a U-shaped function of the sampling fraction, (b) variance in the with-replacement case is a decreasing function of the sampling fraction, and (c) a sampling fraction of $\frac{1}{2}$ in the case of without-replacement bagging produces almost the same variance as a sampling fraction of 1 in the with-replacement case.

3.2 Piecewise-constant target

This represents a situation in which regression trees have the potential to be unbiased because their approximations $\hat{f}(\mathbf{x})$ are piecewise-constant. The target $f(\mathbf{x})$ has the value 1 in the upper corner of a five-dimensional hyper-cube, comprising half of its volume, and the value 0 elsewhere. Thus, the responses $\{y_i\}_1^n$ are roughly evenly divided between those that have $Ey_i = 1$, and those with $Ey_i = 0$. A six terminal node regression tree with five optimally placed splits can exactly represent the target.

Figure 2 shows the bias-squared (3.2) (left frame) and variance (3.3) (right frame) of the without-replacement bagged estimate $\hat{f}(\mathbf{x})$ as a function of the sampling fraction. These values are reported in units of the global target variance $E_{\mathbf{x}}\{f(\mathbf{x}) - E_{\mathbf{x}}f(\mathbf{x})\}^2$. Training samples of $n = 5000$ were used. Although the target lies within the space of the approximating functions, one sees a small bias-squared that decreases with increasing sampling fraction. Unbagged trees have a bias-squared of 0.0025 and variance 0.0863. This bias is due to the fact that the smaller training samples limit the size of the induced trees. This is more dramatically illustrated for smaller training samples. Figure 3 shows the corresponding results for $n = 500$. Here unbagged

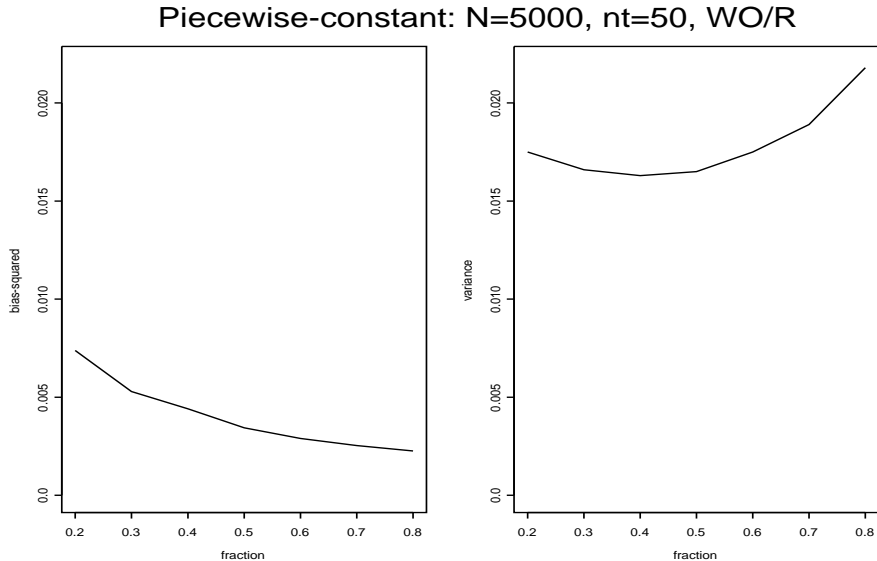


Figure 2: Average bias-squared (left frame) and variance (right frame) as a function of sampling fraction m/n , for without replacement sampling and $n = 5000$, with a piecewise-constant target. Unbagged trees have a bias-squared of 0.0025 and variance 0.0863.

trees had a bias-squared of 0.0325 and variance 0.2440. For the bagged trees, the bias-squared is seen to be comparable to the variance, and the latter increases monotonically with the sampling fraction. This monotonic increase of variance with sampling fraction also occurs for the constant target with small training samples (not shown). This effect is also evident with the larger $n = 5000$ sample (Figure 2) in that the variance is minimized for smaller fractions than with the constant target (Figure 1, left frame). Thus, with bagging there can be a bias-variance trade-off in choosing the sampling fraction m/n . As with any “meta”-parameter that controls bias-variance trade-off, an optimal value can be estimated by minimising an estimate of prediction error through cross-validation or a left out “test” sample.

For completeness Figures 4 and 5 show the corresponding results for $n = 5000$ and $n = 500$ respectively, sampling with replacement. One sees results similar to those for sampling without replacement in the interval $m/n \in [0.2, 0.5]$ of the latter.

The variance plots in Figures 4 and 5 again lend support to the theoretical conclusions reached in Section 2. In particular, properties (a)–(c) noted in the last paragraph of Section 3.1 apply here, too, except that in the case of with-replacement bagging, variance is not a decreasing function of sampling fraction when the latter is large. Note, however, that in the case $n = 500$, shown in Figure 5, variance is an increasing function of sampling fraction, but that by increasing sample size to the “more asymptotic” value $n = 5000$ (Figure 4) the function has almost, but not quite, turned around.

3.3 Linear target

A linear function represents one of the most difficult targets for approximation by regression trees. It does not lie within the space of piecewise-constant functions and its contours are everywhere oblique to the coordinate axes. Figure 6 shows the bias-squared (3.2) (left frame) and variance (3.3) (right frame) of the without-replacement bagged estimate $\hat{f}(\mathbf{x})$ as a function of the sampling fraction for $n = 5000$, again reported in units of the global target variance. Unbagged trees have a bias-squared of 0.0402 and variance 0.2494. Here one sees the dramatic

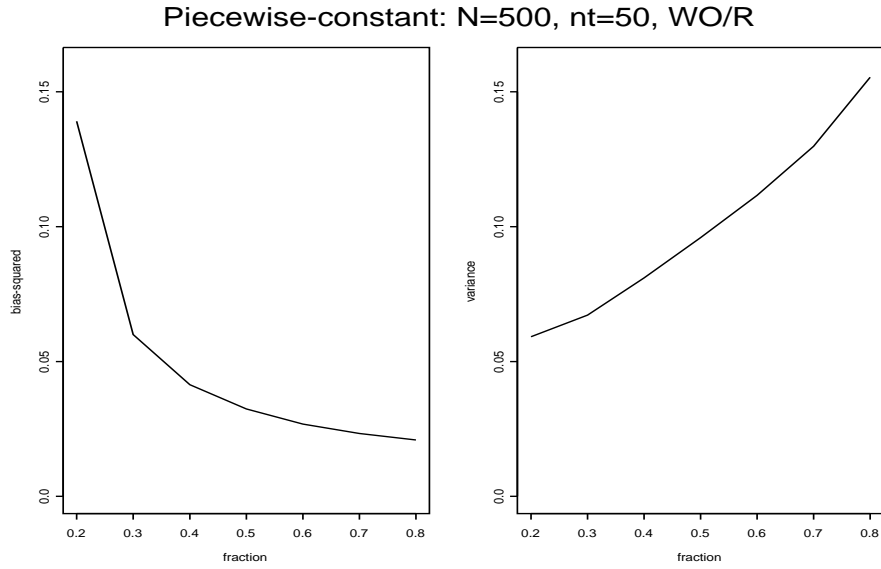


Figure 3: Average bias-squared (left frame) and variance (right frame) as a function of sampling fraction m/n , for without replacement sampling and $n = 500$, with a piecewise-constant target. Unbagged trees have a bias-squared of 0.0325 and variance 0.2440.

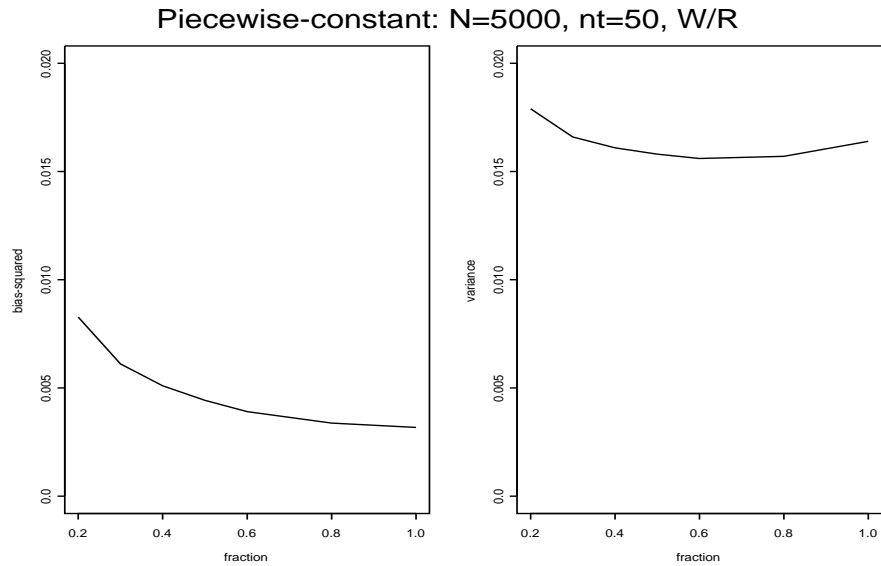


Figure 4: Average bias-squared (left frame) and variance (right frame) as a function of sampling fraction m/n , for with replacement sampling and $n = 5000$, with a piecewise-constant target. Unbagged trees have a bias-squared of 0.0025 and variance 0.0863.

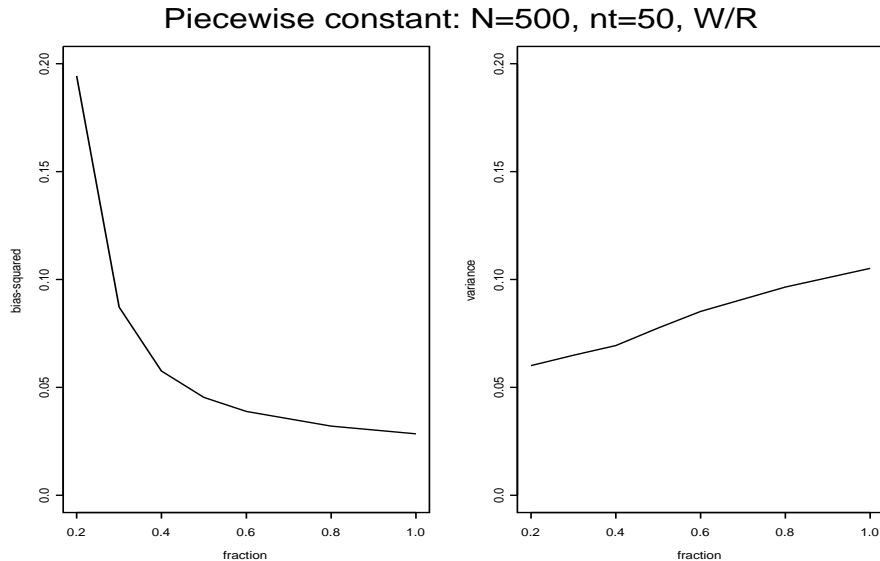


Figure 5: Average bias-squared (left frame) and variance (right frame) as a function of sampling fraction m/n , for with replacement sampling and $n = 500$, with a piecewise-constant target. Unbagged trees have a bias-squared of 0.0325 and variance 0.2440.

super-linear increase of variance with sampling fraction characteristic of the *smaller* ($n = 500$) training samples above. Also, the bias-squared *increases* with sampling fraction. Here bagging is reducing *bias-squared* as well as variance, with smaller sampling fractions producing the most improvement in both. Figure 7 shows the corresponding plot for sampling with replacement. Again the results are similar to the left half (fraction $m/n \leq 0.5$) of the without replacement results.

Figure 8 shows without-replacement results for much larger training samples $n = 50000$. Here unbagged trees have an average bias-squared of 0.0775 and average variance of 0.0821. Comparing to the corresponding (unbagged) $n = 5000$ results above, one sees that using the larger training sample reduces variance, but by a factor of about $1/\sqrt{10}$. However, the bias-squared has *increased* by almost a factor of two. The dependence of bias-squared and variance on sampling fraction is similar to that for $n = 5000$ shown in Fig. 6; they both decrease with decreasing sampling fraction m/n . However, here the bias-squared dominates mean-squared error of the bagged trees.

Although perhaps counter intuitive, the increase in bias-squared with larger samples for fixed-sized regression trees is easy to understand. Figure 9 illustrates the concept in an idealized setting. Shown is a hypothetical asymptotic ($n = \infty$) likelihood as a function of a parameter θ with maximum at $\theta_0 = 0$. Suppose an estimator $\hat{\theta}$ that can realize a discrete set of values $\hat{\theta} \in \{\hat{\theta}_i\} = \{-0.8, -0.3, 0.2, 0.7\}$ (hash marks), none of which is equal to the population parameter value $\theta_0 = 0$. In the asymptotic limit the value of the estimate is $\hat{\theta}_\infty = 0.2$, with a bias-squared of 0.04. However, with finite samples of size n , the expected value of $\hat{\theta}$ is

$$E_n \hat{\theta} = \sum_{i=1}^4 p_i(n) \hat{\theta}_i \quad (3.4)$$

with $p_i(n)$ being the probability that a likelihood based on a sample of size n is maximized at $\hat{\theta}_i$. The dispersion of the $p_i(n)$ -values will tend to increase with decreasing n . This enlarges the set of values that can be realized by (3.4), enabling it to achieve values closer to θ_0 , thereby potentially decreasing the bias.

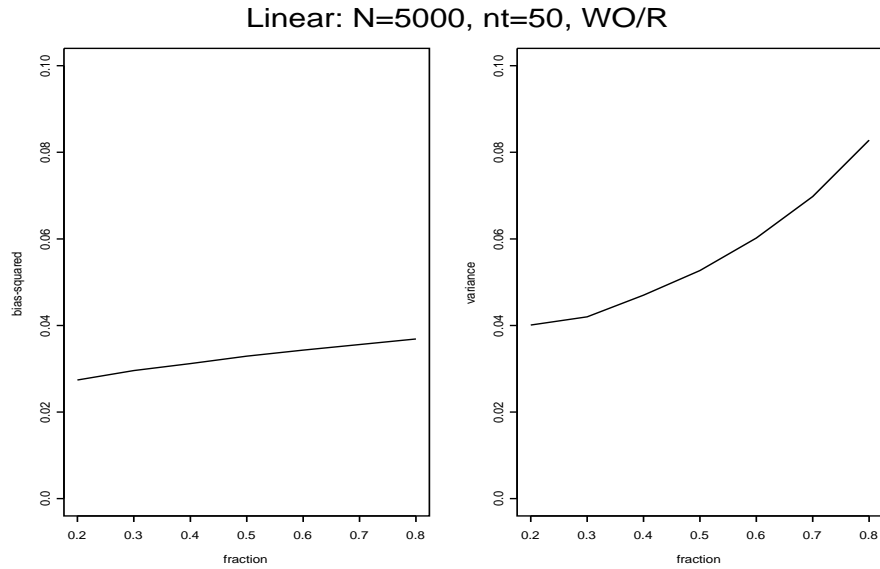


Figure 6: Average bias-squared (left frame) and variance (right frame) as a function of sampling fraction m/n , for without replacement sampling and $n = 5000$, with a linear target. Unbagged trees have a bias-squared of 0.0402 and variance 0.2494.

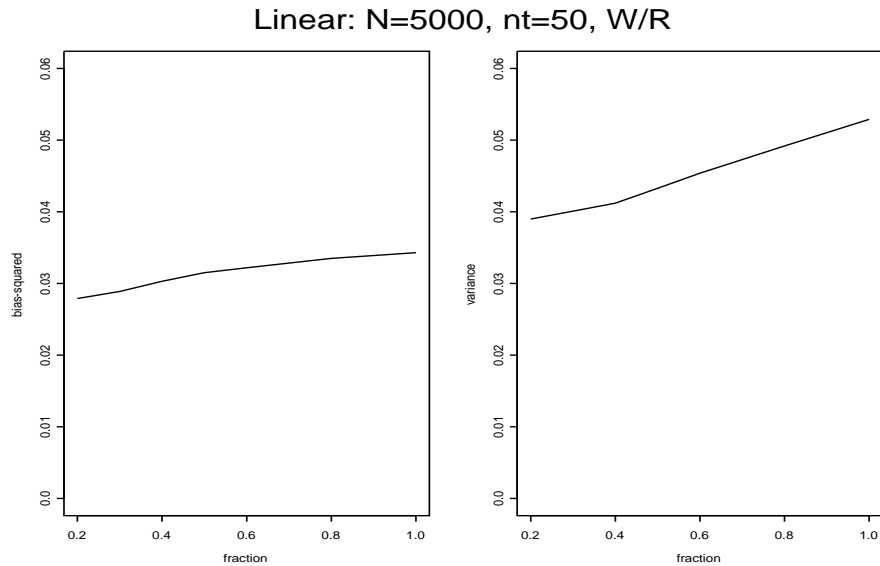


Figure 7: Average bias-squared (left frame) and variance (right frame) as a function of sampling fraction m/n , for with replacement sampling and $n = 5000$, with a linear target. Unbagged trees have a bias-squared of 0.0402 and variance 0.2494.

Linear: N=50000, nt=50, WO/R

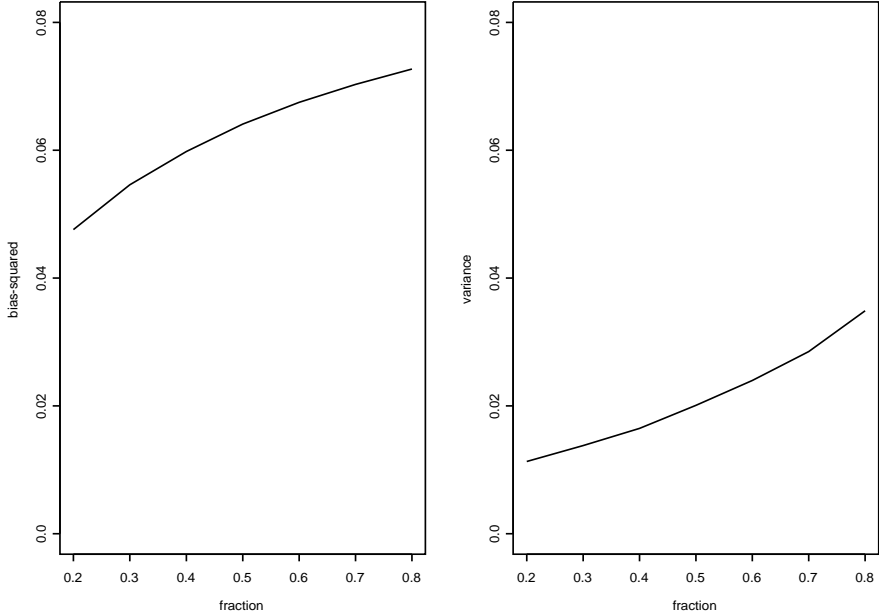


Figure 8: Average bias-squared (left frame) and variance (right frame) as a function of sampling fraction m/n , for without replacement sampling and $n = 50000$, with a linear target. Unbagged trees have an average bias-squared of 0.0775 and average variance of 0.0821.

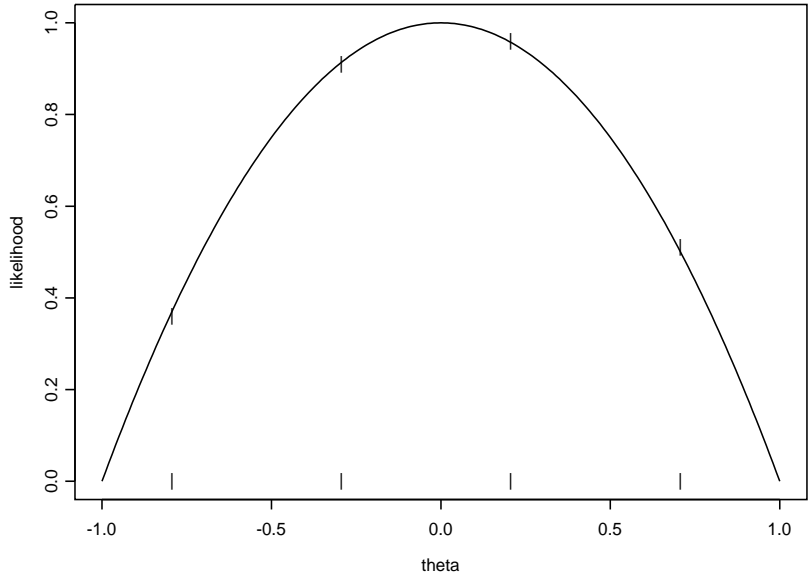


Figure 9: Hypothetical asymptotic likelihood and an estimator that can only realize a discrete set of values.

An L -terminal node regression tree $\hat{f}_L(\mathbf{x})$ cannot get arbitrarily close to a continuous function $f(\mathbf{x})$ such as the linear target used here. Thus, there is an asymptotic average bias-squared characteristic of the closest possible L -terminal node tree. As the sample size is reduced, the distance to the target of the expected approximation $f_L(\mathbf{x}) = E_n \hat{f}_L(\mathbf{x})$ becomes smaller, reducing bias-squared. Of course, this expected approximation $f_L(\mathbf{x})$ is itself not realizable as a (finite sized) regression tree. Bagging reduces bias-squared in such situations simply by reducing the sample size; each bagged tree is trained on a subset of the complete training set (2.7). The averaging aspect of bagging has no effect on bias-squared, but sharply reduces the nonlinear component of variance, thereby producing the win-win situation observed here. However, this argument does not explain why variance increases monotonically with increasing sampling fraction. It may be that in this situation the linear component of the estimator is so small that it does not play a significant role. In such cases the theory of Section 2 cannot provide much insight.

3.4 Bootstrap versus half-sampling

One of the results of the theory is that half-sampling ($m = n/2$) without replacement should produce similar results to full ($m = n$) bootstrap sampling with replacement. In the case of variance, confirmation of this property can be deduced directly from the Figures. Table 1 shows that the relationship also extends to bias-squared, and thus to root mean squared error. The first column identifies each example by the figure(s) in which it was presented. The second and third columns gives the corresponding root-mean-squared estimation error ($\sqrt{\text{bias-squared} + \text{variance}}$) for half-sampling without replacement and full sampling with replacement respectively. The results are seen to verify the theory in this respect.

Table 1

Fig.	$n/2$: W/O	n : W
1	0.100	0.097
2, 4	0.140	0.139
3, 5	0.356	0.364
6, 7	0.292	0.295

Half-sampling of course has a computational advantage, especially if the implementation of the estimator does not support observation weights. More generally, one can see by comparing the relevant figures that m -out-of- n with, and $m/2$ -out-of- n without, replacement sampling give fairly similar results.

4 Acknowledgments

The work of Jerome H. Friedman was partially supported by CSIRO Mathematical and Information Sciences, Australia, the Department of Energy under contract DE-AC03-76SF00515, and by grant DMS9764431 of the National Science Foundation. We're grateful to Rob Tibshirani and two anonymous reviewers for helpful comments.

References

- [1] BREIMAN, L., FRIEDMAN, J.H., OLSHEN, R.A. AND STONE, C.J. (1984). *Classification and Regression Trees*. Wadsworth, Belmont, CA.
- [2] BREIMAN, L. (1996). Bagging predictors. *Machine Learning* **24**, 123–140.
- [3] BREIMAN, L. (1999). Using adaptive bagging to debias regressions. Technical Report No. 547, Department of Statistics, University of California, Berkeley.

- [4] EFRON, B. AND STEIN, C. (1981). The jackknife estimate of variance. *Ann. Statist.* **9**, 586–595.
- [5] EFRON, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*. SIAM, Philadelphia.
- [6] HALL, P. AND SIMAR, L. (1999). Estimating a changepoint, boundary or frontier in the presence of observation error. Manuscript.
- [7] HARTIGAN, J.A. (1969). Using subsample values as typical values. *J. Amer. Statist. Assoc.* **64**, 1303–1317.
- [8] HARTIGAN, J.A. (1971). Error analysis by replaced samples. *J. Roy. Statist. Soc. Ser. B* **33**, 98–110.
- [9] MAHALANOBIS, P.C. (1946). Report on the Bihar crop survey: Rabi season 1943–1944. *Sankhyā* **7**, 269–280.
- [10] MCCARTHY, P.J. (1966). Replication (an approach to the analysis of data from complex surveys). *National Center for Health Statistics*