Asymptotically Optimal Regression Trees

Mohlin, Erik

2018

Document Version:
Other version

Working Paper 2018:12

Department of Economics
School of Economics and Management

Asymptotically Optimal Regression Trees

Erik Mohlin

May 2018

LUND UNIVERSITY

Asymptotically Optimal Regression Trees*

Erik Mohlin†
Lund University

This version May 22, 2018. First version May 5, 2014.

Abstract

Regression trees are evaluated with respect to mean square error (MSE), mean integrated square error (MISE), and integrated squared error (ISE), as the size of the training sample goes to infinity. The asymptotically MSE- and MISE minimizing (locally adaptive) regression trees are characterized. Under an optimal tree, MSE is $O(n^{-2/3})$. The estimator is shown to be asymptotically normally distributed. An estimator for ISE is also proposed, which may be used as a complement to cross-validation in the pruning of trees.

Keywords: Piece-Wise Linear Regression; Partitioning Estimators; Non-Parametric Regression; Categorization; Partition; Prediction Trees; Decision Trees; Regression Trees; Regressogram; Mean Squared Error.
JEL codes: C14; C38.

1 Introduction

Regression trees are an important and widely used tool in machine learning. They are easy to interpret and relatively fast to construct. Moreover they are resistant to the inclusion of irrelevant predictor variables, and are able to handle non-smooth regression surfaces (Hastie et al. 2009, chapter 10). The main drawback is that, for any particular data set, there is usually a more accurate method available. This paper contributes to the theoretical understanding of regression trees, and their predictive accuracy, by analysing mean square error (MSE), mean integrated square error (MISE), and integrated squared error (ISE), as the size of the training sample goes to infinity. The asymptotically MSE-

*This paper has benefited from comments by Liang Chen, Toru Kitagawa, and Bent Nielsen. I am grateful for financial support from the Swedish Research Council (Grant 2015-01751), and the Knut and Alice Wallenberg Foundation (Wallenberg Academy Fellowship 2016-0156).

†Address: Department of Economics, Lund University, Tycho Brahes väg 1, 220 07 Lund, Sweden. E-mail: erik.mohlin@nek.lu.se

and MISE-minimising regression trees are characterized. Asymptotic normality is proved. Moreover, an estimator for ISE is proposed.

A regression tree is a tool for generating piece-wise constant regressions, and as such it can be seen as a particular method for constructing regressograms, introduced by Tukey (1947), and later developed and studied under the name of partitioning estimators (see Györfi et al. (2002) and references therein). Regression trees were introduced by Morgan and Sonquist (1963), and further developed by Breiman et al. (1984), Quinlan (1992), and Loh (2002), among others. Loh (2011) is a recent survey. By a successive procedure of binary splits, a set $\mathcal{X} \subseteq \mathbb{R}^d$ is split into a number of cells in the shape of hyper-rectangles. Together the cells partition the set $\mathcal{X}$. In each step of the splitting procedure, one dimension $\mathcal{X}_j \subseteq \mathbb{R}$ and one splitting point $z \in \mathcal{X}_j$ is used to divide $\mathcal{X}_j$ into two halves. Thus the sequence of splits can be represented as a binary tree τ. The prediction for each cell is equal to the sample mean of the data in that cell. The criterion for evaluating a tree is usually based on the residual sum of squares $R(\tau)$, with the addition of a complexity cost α per split k. First one grows a large tree based on minimisation of $R(\tau)$, and then one reduces the number of cells/splits by pruning. Typically pruning is based on minimisation of $R_\alpha(\tau) = R(\tau) + \alpha k$, where the value of α is determined by some cross-validation procedure. In contrast the current paper studies the bias-variance trade-off analytically. The estimator of ISE that is proposed below may be used as a complement to cross-validation in the pruning of trees.

Decreasing the size of cells has two effects: The within-cell differences between objects tend to decrease, but the number of training observations in each cell tends to decrease, thereby making inference less reliable. It follows that as the size of the training set n is increased, the optimal number of cells is also increased, but at a slower rate, $O\left(n^{-1/3}\right)$. Asymptotically the width of the cell to which x belongs, is (i) increasing in the variance of y conditional on x, (ii) decreasing in the derivative of the mean of y conditional on x, and (iii) decreasing in the marginal density at x. Under an optimal tree, MSE is $O(n^{2/3})$.

In the literature on regression trees the main focus has been on developing algorithms rather than deriving analytical results. Consequently there are few asymptotic results and no results on adaptive optimal cells. Similarly, Györfi et al. (2002) collect many results on consistency of partitioning estimators, but it does not seem to exist any results regarding locally adaptive optimal partitionings. The most closely related results are due to Cattaneo and Farrell (2013). They analyse the asymptotic mean square error of partitionings but do not allow the cells be locally adaptive.[1]

My results on optimal regression trees can be viewed as lying in between those previously obtained for, on the one hand, asymptotically optimal locally adaptive kernels for non-parametric regression (e.g. Fan and Gijbels 1992), and on the other hand, asymp-

[1] When their theorem 3 is restricted to the case of estimation of the regression line (conditional mean) by means of a constant fit, then it makes essentially the same statement about the mean square error as my theorem 1 when all cells are restricted to have the same shape and volume.

totically optimal locally adaptive histograms for non-parametric density estimation (e.g. Kogure 1987). The results partly generalize Mohlin (2014*b*).

2 Model

2.1 Data

Consider a sample, or training set, $\{(X^s, Y^s)\}_{s=1}^n$ of size n. Each realised observation is a $d+1$-dimensional attribute vector $(x^s, y^s) = (x_1^s, x_2^s, ...x_d^s, y^s) \in \mathcal{X} \times \mathcal{Y}$, where $\mathcal{Y} = \mathbb{R}$, and $\mathcal{X} = \times_{j=1}^d \mathcal{X}_j$ is a compact subset of $\mathbb{R}^n$, such that $\mathcal{X}_j = [a_j, b_j] \subseteq \mathbb{R}$ for all dimensions j. Observations are i.i.d. draws from a probability distribution with joint density $f(x, y)$ and marginal densities $f_X(x)$ and $f_Y(y)$.[2] The density of Y conditional on $X = x$ is $f_Y(y|x) = f(x, y)/f_X(x)$, assuming $f_X(x) > 0$ for all $x \in \mathcal{X}$. For all $X \in \mathcal{X}$, write

$$Y = m(X) + \varepsilon(X),$$

where $m(X) = \mathbb{E}[Y|X]$ is the conditional mean, and $\varepsilon(X) = Y - \mathbb{E}[Y|X]$ is a noise term with mean zero. Assume $|\mathbb{E}[Y|X = x]| < \infty$ so that the variance $\sigma^2(x) = Var(Y|X = x)$ is well defined.

The set of samples, or training sets, of size n is denoted $\mathcal{T}(n)$.

2.2 Trees

A tree τ induces a set of cells (or categories) $\mathcal{C} = \{\mathcal{C}^1, ..., \mathcal{C}^k\}$ that partitions $\mathcal{X}$. Each cell is a hyper-rectangle i.e. $\mathcal{C}^i = \times_{j=1}^d \mathcal{C}_j^i$, where component $\mathcal{C}_j^i$ is an interval (open, closed or half-open) of length h_j^i on $\mathbb{R}$, such that

$$\inf_{x_j} \mathcal{C}_j^i = a_j^i, \qquad \sup_{x_j} \mathcal{C}_j^i = a_j^i + h_j^i = b_j^i.$$

We identify a tree τ by the set of cells (or categories) $\mathcal{C}$ that it induces. The probability that an object belongs to cell/category i is $p^i = \int_{x \in \mathcal{C}^i} f_X(x)\, dx$, which is required to be strictly positive. The conditional marginal density of y in $\mathcal{C}^i$ is $f_Y(y|x \in \mathcal{C}^i) = \int_{x \in \mathcal{C}^i} f(x, y)\, dx/p^i$, and the conditional marginal density of x in $\mathcal{C}^i$ is $f_X(x|x \in \mathcal{C}^i) = f_X(x)/p^i$. Denote the within-category variance $Var(Y^i) = Var(Y|X \in \mathcal{C}^i)$, and the within-category mean $\mu^i = \mathbb{E}[m(X)|X \in \mathcal{C}^i]$.

The relative size of cells is constrained by some (small) number $\rho \in (0, 1)$ such that, for all i and j, if $p^i \geq p^j$ then $p^j \geq \rho p^i$. For any finite number of cells this simply means

[2] More formally observations are drawn i.i.d. according to an absolutely continuous cumulative distribution function $F : \mathcal{V} \to [0, 1]$, with a bounded probability density function, $f : \mathcal{V} \to \mathbb{R}_+$. The marginal densities are $f_X(x) = \int_{y \in \mathcal{Y}} f(x, y)\, dy$ and $f_Y(y) = \int_{x \in \mathcal{X}} f(x, y)\, dx$.

that all cells have positive probability. When the number of cells goes to infinity, the constraint implies that no cell becomes relatively infinitely larger than another cell. The set of categorizations satisfying the assumptions above, called feasible categorizations, is denoted Ψ. For a given training set t the set of feasible categorizations with non-empty cells is $\Psi(t)$.

2.3 Prediction Error and Optimality

Given a training set t, each cell $\mathcal{C}^i$ is associated with a unique (point) prediction $\hat{y}^i$. Let $\hat{y}(x)$ be the prediction for the cell to which x belongs. The *squared error (SE)* associated with an object (x, y) is

$$SE(\mathcal{C}, t)(x, y) = (y - \hat{y}(x))^2. \tag{1}$$

In the statistical learning literature this is also known as the prediction error. Taking expectation over objects (X, Y) we obtain the *integrated squared error (ISE)*, which is also known as generalization error, or test error,

$$ISE(\mathcal{C}, t) = \mathbb{E}\left[(Y - \hat{y}(X))^2\right] = \int \mathbb{E}\left[(Y(x) - \hat{y}(x))^2\right] f_X(x)\, dx. \tag{2}$$

Fixing $X = x$ and taking expectation over Y one may define a point-wise version of *ISE*, $ISE(\mathcal{C}, t)(x) = \mathbb{E}\left[(Y - \hat{y}(X))^2 | X = x\right]$.

Ex ante the training set is a random variable T, and hence the resulting prediction is also a random variable $\hat{Y}$. Fixing $X = x$ and taking the expectation of $SE(\mathcal{C}, t)(x)$ over $\hat{Y}$ and Y yields the *mean square error (MSE)*

$$MSE(\mathcal{C}, n)(x) = \mathbb{E}\left[\left(Y(x) - \hat{Y}(x)\right)^2\right] = \mathbb{E}\left[\left(Y - \hat{Y}(X)\right)^2 | X = x\right]. \tag{3}$$

Now taking expectation over X yields the *mean integrated squared error (MISE)*

$$MISE(\mathcal{C}, n) = \mathbb{E}\left[\left(Y - \hat{Y}(X)\right)^2\right] = \int \mathbb{E}\left[\left(Y(x) - \hat{Y}(x)\right)^2\right] f_X(x)\, dx. \tag{4}$$

In the statistical learning literature *(MSE)* and *(MISE)* are also known as the expected prediction error or expected test error.

Remark 1 *Some authors (e.g. Härdle 1990) focus on conditional square error* $(m(x) - \hat{y}(x))$*, and hence on conditional MSE and conditional MISE, thereby ignoring the irreducible error* $\sigma^2(x)$*. Results would essentially be unaffected by using* $(m(x) - \hat{y}(x))^2$ *instead of* $(y - \hat{y}(x))^2$ *as the basis of evaluation. In particular the solution derived in theorems 1 and 2 would be the same.*

2.4 Prediction

The prediction rule remains to be specified. Fixing the training set one may look for predictions $\{\hat{y}^i\}_{i=1}^k$ that minimize $ISE(\mathcal{C}, t)$, for a given categorization $\mathcal{C}$. Thus, for each cell i the optimal prediction $\hat{y}^i$ solves

$$\min_{\hat{y}^i} \mathbb{E}\left[\left(Y - \hat{y}^i\right)^2 | X \in \mathcal{C}^i\right]. \tag{5}$$

It is basic result of statistical decision theory that mean square error is minimized when the prediction is equal to the conditional mean (see e.g. theorem 2.1 of Li and Racine 2007). In the current set-up this means that the solution to (5) is $\hat{y}^i = \mathbb{E}[Y|X \in \mathcal{C}^i] = \mu^i$. The true mean μ^i is not observed, but the sample mean $\bar{y}^i$ is an unbiased estimator of the true mean. For this reason the within-cell mean $\bar{y}^i$ will be used as prediction, as long as the cell is non-empty. The choice of predictor for empty cells will turn out to be unimportant for the results. For simplicity I assume that the prediction for an empty cell is the mean across all observations, $\bar{y}$. The training set for cell i is $t^i = t \cap \mathcal{C}^i$. Let $n^i = |t^i|$, so that $\sum_{i=1}^k n^i = n$. The prediction for cell i is

$$\hat{y}^i = \begin{cases} \bar{y}^i = \frac{1}{n^i} \sum_{y^s \in t_i} y^s & \text{if } n^i > 0 \\ \bar{y} = \frac{1}{n} \sum_{s=1}^n y^s & \text{if } n^i = 0 \end{cases}. \tag{6}$$

Note that ex ante the number of observations in cell i is a random variable N^i.

3 Results

3.1 Preliminary Results

The first two results provide expressions for *ISE*, *MSE* and *MISE*.

Lemma 1 *The* integrated square error *for a categorization $\mathcal{C} \in \Psi(t)$, conditional on a training set t, is*[3]

$$ISE(\mathcal{C}, t) = \sum_{i=1}^k p^i \left(Var\left(Y^i\right) + \left(\bar{y}^i - \mu^i\right)^2\right).$$

Lemma 2 *The* mean squared error *of categorization $\mathcal{C} \in \Psi$, at $x \in \mathcal{C}^i$, is*

$$MSE(\mathcal{C}, n)(x) = \sigma^2(x) + \left(Bias\left(\bar{Y}^i\right)\right)^2 + Var\left(\bar{Y}^i\right),$$

where

$$\left(Bias\left(\bar{Y}^i\right)\right)^2 = \left(m(x) - \mu^i\right)^2,$$

[3] Point-wise one may also calculate, $ISE(\mathcal{C}, t)(x) = \sigma^2(x) + (m(x) - \bar{y}_{it})^2$.

and

$$Var\left(\bar{Y}^i\right) = Var\left(Y^i\right)\sum_{r=1}^{n}\Pr\left(N^i = r\right)\frac{1}{r} + \Pr\left(N^i = 0\right)\mathbb{E}\left[\left(\bar{Y}_n - \mu^i\right)^2 | N^i = 0\right],$$

with $N^i \sim Bin\left(n, p^i\right)$.

Lemma 3 τ*The* mean integrated squared error *is*

$$\begin{aligned} MISE\left(\mathcal{C}, n\right) &= \sum_{i=1}^{k} p^i\left(\mathbb{E}\left[\sigma^2\left(X\right)|X \in \mathcal{C}^i\right] + Var\left(m\left(X\right)|X \in \mathcal{C}^i\right) + Var\left(\bar{Y}^i\right)\right) \\ &= \sum_{i=1}^{k} p^i\left(Var\left(Y^i\right) + Var\left(\bar{Y}^i\right)\right). \end{aligned}$$

The expression for $MSE\left(\mathcal{C}, n\right)(x)$ reveals the fundamental bias-variance trade-off. In addition the prediction error is affected by *irreducible noise* $\sigma^2\left(x\right)$.

Remark 2 *It is easy to see that* ISE, MSE, *and* $MISE$ *are continuous in the distribution* f. *Mohlin (2014*a*) defines a metric on partitions which can be employed to show that prediction errors* ISE, MSE, *and* $MISE$ *are also continuous in the decision variable* $\mathcal{C}$.

3.2 *MSE*- and *MISE*-Optimal Trees

Intuitively, as n increases one may obtain a reasonable approximation of $m\left(x\right)$ by increasing the number of cells k, though at a slower rate than n. The following lemma formalizes this intuition:

Lemma 4 *For any* $\varepsilon > 0$ *there are* $n_0 > 0$ *and* $\delta > 0$, *such that if* $n > n_0$ *and*

$$MISE\left(\mathcal{C}, n\right) - \inf_{\mathcal{C}' \in \Psi} MISE\left(\mathcal{C}', n\right) < \delta,$$

then $\mathcal{C}$ *satisfies* $1/k < \varepsilon$, *and* $k/n < \varepsilon$.

The lemma says that minimization of $MISE\left(\mathcal{C}, n\right)$ implies that if the number of observations goes to infinity ($n \to \infty$) then it is optimal to let the number of cells go to infinity too ($k \to \infty$), but at a slower rate ($k/n \to 0$). This is the basis for the asymptotic results that will now be derived.

3.2.1 Fixed Design

Consider first a fixed design model where data is taken such that $n^i = np^i$. (See e.g. Jennen-Steinmetz and Gasser (1988) for a similar fixed design assumption.) If f is three times differentiable, Taylor approximations may be used to derive the asymptotically optimal width of cells as $n \to \infty$.

As we let $n \to \infty$ there exist corresponding sequences of optimal trees (for each n there is at least one optimal tree τ^n). For a given sequence of optimal trees $\{\tau^n\}_{n=1}^{\infty}$ we examine the sequence of cells $\left\{\mathcal{C}^{i(n)}(x)\right\}_{n=1}^{\infty}$ that contain an arbitrary point x.

Theorem 1 *Let $\{\tau^n\}_{n=1}^{\infty}$ be a sequence of optimal trees, and let $\left\{\mathcal{C}^{i(n)}(x)\right\}_{n=1}^{\infty}$ be the sequence of cells that contain x. Supressing the dependence on n, let the volume of $\mathcal{C}^{i(n)}(x)$ be denoted $h = \times_{l=1}^{d} h_l$. Suppose that f is three times differentiable at x, and $f_X(x) > 0$. Assume a fixed design model such that $n^i = np^i$. If $n \to \infty$, $k \to \infty$, and $k/n \to 0$ (implying $p^i \to 0$ and $np^i \to \infty$), then asymptotically,*

$$MSE(\mathcal{C}, n)(x) = \sigma^2(x) + \underbrace{h^2\left(\sum_j \frac{\partial m(x)}{\partial x_j}\delta_j\right)^2 + O\left(h^3\right)}_{Bias\left[\bar{Y}(x)\right]^2} + \underbrace{\frac{\sigma^2(x)}{f_X(x)\, nh} + O\left(n^{-1}\right)}_{Var\left[\bar{Y}(x)\right]}, \quad (7)$$

where $\delta_j \in [0,1]$ is such that $\left|\mathbb{E}\left[s_j | s_j \in \mathcal{C}_j^{i(n)}\right] - x_j\right| = \delta_j h_j$. If $\sum_j \frac{\partial m(x)}{\partial x_j}\delta_j \neq 0$ then the asymptotic $MSE(\mathcal{C}, n)(x)$ is minimized by

$$h = \left(\frac{\sigma^2(x)}{2n f_X(x)\left(\sum_j \frac{\partial m(x)}{\partial x_j}\delta_j\right)^2}\right)^{\frac{1}{3}}, \quad (8)$$

which yields

$$MSE(\mathcal{C}', n)(x) = \sigma^2(x) + O\left(n^{-\frac{2}{3}}\right). \quad (9)$$

The size of cells should decrease (and the number of cells should increase) at a rate $O\left(n^{-d/3}\right)$. Asymptotically, the size of the MSE-minimising cell is decreasing in the density $f_X(x)$, and the curvature $\sum_j \frac{\partial m(x)}{\partial x_j}$ of the conditional mean. It is increasing in the variance $\sigma^2(x)$.

Remark 3 *If f is not continuously differentiable but Lipschitz continuous we may still derive an upper bound on the mean square error, and an expression for the associated width h^*. Restrict attention to $d = 1$ and a fixed design model with $n^i = np^i$. Assume*

that $\sigma^2(x)$, $m(x)$, *and* $f(x)$ *are Lipschitz, so that for any* $x, \alpha, \beta, \gamma \in \mathcal{C}^i$ *it holds that*

$$\begin{aligned}\left|\sigma^2(x) - \sigma^2(\alpha)\right| &< \lambda |x - \alpha|, \\ |m(x) - m(\beta)| &< \eta |x - \beta|, \\ |f_X(x) - f(\gamma)| &< \phi |x - \gamma|.\end{aligned}$$

In this case we find that

$$MSE(C, n)(x) < \sigma^2(x) + \eta^2 h^2 + \frac{\sigma^2(x)}{f(x) hn} + O\left(n^{-1}\right).$$

The right hand side is minimised by

$$h = \left(\frac{\sigma^2(x)}{2n\eta^2 f(x)}\right)^{\frac{1}{3}}.$$

Evaluating $MSE(C, n)(x)$ *with this solution we find*

$$MSE(C, n)(x) = \sigma^2(x) + O\left(n^{-\frac{2}{3}}\right).$$

The results are is similar to those obtained with Taylor approximations. Note that η *is a measure of the roughness of the conditional mean, just like* m'.

3.2.2 Random Design

We now verify that the speed of convergence is the same under a random design model as under the fixed design model.

Theorem 2 *Let* $\{\tau^n\}_{n=1}^{\infty}$ *be a sequence of optimal trees, and let* $\left\{\mathcal{C}^{i(n)}(x)\right\}_{n=1}^{\infty}$ *be the sequence of cells that contain* x. *Supressing the dependence on* n, *let the volume of* $\mathcal{C}^{i(n)}(x)$ *be denoted* $h = \times_{l=1}^{d} h_l$. *Suppose that* f *is three times differentiable at* x, *and* $f_X(x) > 0$. *Assume a random design model. If* $n \to \infty$, $k \to \infty$, *and* $k/n \to 0$ *(implying* $p^i \to 0$ *and* $np^i \to \infty$*), then asymptotically,*

$$MSE(\mathcal{C}, n)(x) = \sigma^2(x) + \underbrace{h^2 \left(\sum_j \frac{\partial m(x)}{\partial x_j} \delta_j\right)^2 + O\left(h^3\right)}_{Bias\left[\bar{Y}(x)\right]^2} + \underbrace{\frac{\sigma^2(x)}{f_X(x)} O\left((nh)^{-1}\right) + O\left(n^{-1}\right)}_{Var\left[\bar{Y}(x)\right]}, \tag{10}$$

where $\delta_j \in [0, 1]$ *is such that* $\left|\mathbb{E}\left[s_j | s_j \in \mathcal{C}_j^{i(n)}\right] - x_j\right| = \delta_j h_j$. *If* $\sum_j \frac{\partial m(x)}{\partial x_j} \delta_j \neq 0$ *then the*

asymptotic $MSE\left(\mathcal{C},n\right)\left(x\right)$ *is minimized by*

$$h=O\left(n^{-\frac{1}{3}}\right), \tag{11}$$

which yields

$$MSE\left(\mathcal{C}',n\right)\left(x\right)=\sigma^{2}\left(x\right)+O\left(n^{-\frac{2}{3}}\right). \tag{12}$$

3.2.3 Asymptotic Normality

The regression tree estimator is asymptotically normally distributed. This holds regardless of whether a fixed or random design is assumed.

Theorem 3 *Suppose that* f *is three times differentiable at* $x\in\mathcal{C}^{i}$*, a point in the interior of the support of* X*, and* $f_{X}\left(x\right)>0$*,* $\sigma^{2}\left(x\right)>0$*. If* $n\rightarrow\infty$*,* $k\rightarrow\infty$*, and* $k/n\rightarrow0$*, then asymptotically,*

$$\sqrt{n^{i}}\left[\bar{Y}\left(x\right)-m\left(x\right)-h^{2}\left(\sum_{j}\frac{\partial m\left(x\right)}{\partial x_{j}}\delta_{j}\right)^{2}\right]\xrightarrow{d}N\left(0,\sigma^{2}\left(x\right)\right).$$

3.3 *ISE*-Optimal Trees

Deriving the optimal trees by computing MSE and $MISE$ relies on knowledge of the underlying distribution f. In practice this information is not available to the analyst, so the optimal tree is usually determined via cross-validation. An alternative approach would be to find an estimator of MSE and $MISE$. I would like to suggest the following estimator:

Definition 1 *Let* $\hat{\Psi}\left(t\right)$ *denote the set of feasible trees in which all cells have at least two elements* $(n^{i}\geq2)$ *given the training set* t*. The* sample integrated square error (SISE) *for a tree* $\mathcal{C}\in\hat{\Psi}\left(t\right)$*, conditional on a training set* t*, is*

$$SISE\left(\mathcal{C},t\right)=\sum_{i=1}^{k}\frac{n^{i}}{n}\left(1+\frac{1}{n^{i}}\right)s_{i}^{2},\qquad s_{i}^{2}=\frac{1}{n^{i}-1}\sum_{y^{s}\in t^{i}}\left(y^{s}-\bar{y}^{i}\right)^{2}.$$

Note that this can be written

$$SISE\left(\mathcal{C},t\right)=\sum_{i=1}^{k}\frac{1}{n}\frac{n^{i}+1}{n^{i}-1}\sum_{y^{s}\in t^{i}}\left(y^{s}-\bar{y}^{i}\right)^{2}.$$

In the absence of the factor $\left(n^{i}+1\right)/\left(n^{i}-1\right)$ this expression would simply be the average residual sum of squares (RSS), which is also known as training error in the statistical

learning literature. The factor $(n^i + 1)/(n^i - 1)$ penalises cells with few members. The motivation for this particular choice of adjustment of RSS comes from the following observations:

Theorem 4 *For a given tree τ, inducing categorization $\mathcal{C}$, and training set $\tilde{t}$ with an allocation of observations to cells $\{\tilde{n}^1, \tilde{n}^2, ..., \tilde{n}^k\}$, such that $\tilde{n}^i \geq 2$ for all i, let $\mathcal{T}\left(\mathcal{C}, \tilde{t}\right)$ be the set of training sets t such that $\tilde{n}^i = n^i$ for each cell i in $\mathcal{C}$.*

(a) *If expectation is taken over $\mathcal{T}\left(\mathcal{C}, \tilde{t}\right)$, then*

$$\mathbb{E}\left[ISE\left(\mathcal{C}, t\right)\right] = \sum_{i=1}^{k} p^i \left(1 + \frac{1}{n^i}\right) Var\left(Y^i\right),$$

and

$$\mathbb{E}\left[SISE\left(\mathcal{C}, t\right)\right] = \sum_{i=1}^{k} \frac{n^i}{n} \left(1 + \frac{1}{n^i}\right) Var\left(Y^i\right).$$

(b) *For any sequence of training sets from $\mathcal{T}\left(\mathcal{C}, \tilde{t}\right)$*

$$P \lim_{n\to\infty} ISE\left(\mathcal{C}, t\right) = P \lim_{n\to\infty} SISE\left(\mathcal{C}, t\right) = \sum_{i=1}^{k} p^i Var\left(Y^i\right).$$

Part (a) of the theorem implies that if the actual fraction of objects in each cell, n^i/n, is equal to the probability of receiving an object in the corresponding cell p^i (as it was under the fixed design assumption above), then $ISE\left(\mathcal{C}, t\right)$ and $SISE\left(\mathcal{C}, t\right)$ have the same expected value on $\mathcal{T}\left(\mathcal{C}, t\right)$ (the set of training sets such that $\tilde{n}^i = n^i$ for each cell i in $\mathcal{C}$). Part (b) confirms that for large enough n it is highly probable that $n^i/n = p^i$.

The estimator $SISE\left(\mathcal{C}, t\right)$ could be employed in the pruning of regression trees. (Note that $SISE\left(\mathcal{C}, t\right)$ is well-defined also when the input variables are categorical rather than numerical.) This would not require any use of cross-validation, since the cost-complexity trade-off is decided on analytically. Indeed a $SISE\left(\mathcal{C}, t\right)$-based pruning process might be seen as a complement to traditional cross-validation methods. It should be straightforward to modify the tree algorithm in the statistical software R since the command `prune.tree` has an optional `method` argument whose default `method="deviance"` minimises average RSS.

4 Discussion

Non-parametric kernel regression is another common method for estimating the conditional mean $m\left(x\right)$. To facilitate comparison with the results above restrict attention to

the one-dimensional case, $d = 1$. In this case (11) becomes

$$h_C^*(x) = \left(\frac{\sigma^2(x)}{2n f_X(x) (\delta m'(x))^2} \right)^{1/3}. \tag{13}$$

Fan and Gijbels (1992) derive the following expression for the locally adaptive asymptotically optimal kernel bandwidth (where optimal is understood in the sense of minimizing conditional MISE),

$$h_K^*(x) = q_K \left(\frac{\sigma^2(x)}{n f_X(x) (m''(x))^2} \right)^{1/5}, \tag{14}$$

where q_K is a constant which is independent of x. This is very similar to the expression (13) for the optimal cell width $h_C^*(x)$. One difference is that the curvature of the conditional mean enters through the second derivative $m''(x)$ here, compared to the first derivative $m'(x)$ above. The reason is that the kernel is symmetric around x, whereas the cell induced by a regression tree is not, except for a measure zero set of points. Another difference is that speed at which the bandwidth vanishes is slower than the speed at which the optimal cell width vanishes.

The kernels due to Priestley and Chao (1972), and Gasser and Müller (1979), lead to expressions that are similar to (14), see e.g. Härdle (1990) and Brockmann et al. (1993). The kernel of Nadaraya (1964) and Watson (1964) (being of first order) induces a bias term that involves both the first and second order derivatives of m at x.

Within density estimation histograms is the closest related approach. Kogure (1987) (see also Scott 1992) derives the following expression for the locally adaptive asymptotically optimal bin width,

$$h_H^*(x) = q_H \left(\frac{f_X(x)}{n (f_X'(x))^2} \right)^{1/3}, \tag{15}$$

where q_H is a constant which is independent of x. In contrast to (13) and (14) this expression is independent of the variance $\sigma^2(x)$, and the optimal width at x is increasing in density $f_X(x)$, rather than decreasing. Interestingly, the width of the bins optimally vanishes at the rate $O\left(n^{-1/3}\right)$, exactly as in (13).

In conclusion, the optimality result in theorem 2 lies between the results derived for kernels in regression analysis and histograms in density estimation. This reflects the fact that while the prediction *task* of regression trees and is essentially the same as that of kernel regression, the *tool* (producing an estimate for each cell in a partition) is more similar to histograms.

5 Binary Outcomes / Classification

The framework developed above can be amended to handle the task of estimating the probability of binary outcomes, with the aid of a tree. The binary outcome space could signify membership in one of two classes. However, the question that is being asked is different from the one posed by classification trees (e.g. Ripley (1996)). In that literature the prediction is class membership, not the probability of a particular class membership. Moreover the framework of this paper uses loss functions based on squared error, something that is rarely the criterion for evaluation of classification trees.

5.1 Model

5.1.1 Data and Trees

Instead of a continuous outcome set, there is a binary outcome set $\mathcal{Y} = \{0, 1\}$. Furthermore, instead of making a point prediction of what value Y will take (in the set $\mathcal{Y}$), the task will be to estimate the probability that Y takes either of the values in $\mathcal{Y} = \{0, 1\}$. The density $f(x, y)$, the marginal densities $f_X(x)$ and $f_Y(y)$, and the conditional density $f_Y(y|x)$ are all defined as above. As before, let $m(x) = \mathbb{E}[Y|X = x]$. Note that $\mathbb{E}[Y|X = x] = f_Y(1|x)$. Thus the task is to estimate

$$m(x) = f_Y(1|x) = \mathbb{E}[Y = 1|X = x] \in [0, 1].$$

One may still use the decomposition $Y = m(X) + \varepsilon(X)$.

A tree is defined as above, along with the probability p^i, the conditional marginal densities $f_Y(y|x \in \mathcal{C}^i)$ and $f_X(x|x \in \mathcal{C}^i)$. Feasible categorizations are also defined as before.

5.1.2 Prediction Error

Given a training set t, each cell $\mathcal{C}^i$ is associated with a unique (point) estimate $\hat{m}^i$. Let $\hat{m}(x)$ be the estimate for the cell to which x belongs. Using $\hat{m}(x)$ instead of $\hat{y}(x)$, and $m(x)$ instead of $y(x)$ (and $Y(x)$), in equations (1), (2), (3), and (4), redefine squared error (SE), integrated squared error (ISE), mean square error (MSE), and mean integrated squared error (MISE), for the present set-up.[4] For example

$$ISE'(\mathcal{C}, t) = \mathbb{E}\left[m(X)) - \hat{m}(X)^2\right] = \int \mathbb{E}\left[(m(x) - \hat{m}(x))^2\right] f_X(x)\, dx.$$

[4] This brings the definitions closer to the ones of Härdle (1990), mentioned above.

and

$$MSE'\left(\mathcal{C},n\right)\left(x\right)=\mathbb{E}\left[\left(m\left(x\right)-\hat{M}\left(x\right)\right)^{2}\right]=\mathbb{E}\left[\left(m\left(X\right)-\hat{M}\left(X\right)\right)^{2}|X=x\right].$$

Here $\hat{M}$ reflects the fact that the estimator is itself a random variable, just as the use of $\hat{Y}\left(X\right)$ in equation (3)

5.1.3 Prediction

By the same line of reasoning as before, look for a prediction $\hat{y}^{i}$, for each cell i, that minimize

$$\min_{\hat{y}^{i}}\mathbb{E}\left[\left(m\left(X\right)-\hat{m}^{i}\right)^{2}|X\in\mathcal{C}^{i}\right],$$

and for the same reasons as before the solution is $\hat{m}^{i}=\mathbb{E}\left[m\left(X\right)|X\in\mathcal{C}^{i}\right]=\mu^{i}$. Again, the true mean μ^{i} is not observed, but the sample mean $\bar{y}^{i}$ is an unbiased estimator of it. Consequently the estimator of $\hat{m}^{i}$ should be equal to the predictor $\hat{y}^{i}$ used above:

$$\hat{m}^{i}=\left\{\begin{array}{ll}\bar{y}^{i}=\frac{1}{n^{i}}\sum_{y^{s}\in t_{i}}y^{s} & \text{if } n^{i}>0\\ \bar{y}=\frac{1}{n}\sum_{s=1}^{n}y^{s} & \text{if } n^{i}=0\end{array}\right..$$

5.2 Results

The results are almost identical to the case of a continuous outcome variable. The only difference is that the irreducible noise vanishes. This is due to the fact that Y has been replaced with $m\left(X\right)$ in the definitions of prediction error. I only review the results for (the revised version of) mean square error. Corresponding to lemma 2 the following holds:

Lemma 5 *The* mean squared error *of categorization $\mathcal{C}\in\Psi$, at $x\in\mathcal{C}^{i}$, is*

$$MSE'\left(\mathcal{C},n\right)\left(x\right)=\left(Bias\left(\bar{Y}^{i}\right)\right)^{2}+Var\left(\bar{Y}^{i}\right),$$

where $\left(Bias\left(\bar{Y}^{i}\right)\right)^{2}$ and $Var\left(\bar{Y}^{i}\right)$ are the same as in lemma 2.

In the same way theorem 1 is recovered, the only difference being that the irreducible noise is taken away from equation (7).

Theorem 5 *Let $\left\{\tau^{n}\right\}_{n=1}^{\infty}$ be a sequence of optimal trees, and let $\left\{\mathcal{C}^{i(n)}\left(x\right)\right\}_{n=1}^{\infty}$ be the sequence of cells that contain x. Supressing the dependence on n, let the volume of $\mathcal{C}^{i(n)}\left(x\right)$ be denoted $h=\times_{l=1}^{d}h_{l}$. Suppose that f is three times differentiable at x, and $f_{X}\left(x\right)>0$. Assume a fixed design model such that $n^{i}=np^{i}$. If $n\to\infty$, $k\to\infty$, and $k/n\to 0$*

(implying $p^i \to 0$ and $np^i \to \infty$), then asymptotically,

$$MSE\left(\mathcal{C},n\right)\left(x\right) = \underbrace{h^2\left(\sum_j \frac{\partial m\left(x\right)}{\partial x_j}\delta_j\right)^2 + O\left(h^3\right)}_{Bias\left[\bar{Y}(x)\right]^2} + \underbrace{\frac{\sigma^2\left(x\right)}{f_X\left(x\right)nh} + O\left(n^{-1}\right)}_{Var\left[\bar{Y}(x)\right]},$$

where $\delta_j \in [0,1]$ is such that $\left|\mathbb{E}\left[s_j | s_j \in \mathcal{C}_j^i\right] - x_j\right| = \delta_j h_j$. If $\sum_j \frac{\partial m(x)}{\partial x_j}\delta_j \neq 0$ then the asymptotic $MSE\left(\mathcal{C},n\right)\left(x\right)$ is minimized by (8), which yields (9).

6 Conclusion

Hopefully the theoretical investigations of this paper will somehow prove useful in the further development and application of regression trees and related machine learning methods.

References

Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984), *Classification and Regression Trees*, Wadsworth, New York.

Brockmann, M., Gasser, T. and Herrmann, E. (1993), 'Locally adaptive bandwidth choice for kernel regression estimators', *Journal of the American Statistical Association* **88**(424), 1302–1309.

Cattaneo, M. D. and Farrell, M. H. (2013), 'Optimal convergence rates, bahadur representation, and asymptotic normality of partitioning estimators', *Journal of Econometrics* **174**(2), 127–143.

Chung, F. and Lu, L. (2002), 'Connected components in random graphs with given expected degree sequences', *Annals of Combinatorics* **6**(2), 125–145.

Fan, J. and Gijbels, I. (1992), 'Variable bandwidth and local linear regression smoothers', *The Annals of Statistics* **20**, 2008–2036.

Gasser, T. and Müller, H.-G. (1979), Kernel estimation of regression functions, *in* T. Gasser and M. Rosenblatt, eds, 'Smoothing Techniques for Curve Estimation', Springer Verlag, New York, pp. 23–68.

Györfi, L., Krzyżak, A., Kohler, M. and Walk, H. (2002), *A distribution-free theory of nonparametric regression*, Springer.

Hastie, T., Tibshirani, R. and Friedman, J. (2009), *The Elements of Statistical Learning Theory*, Springer, New York.

Härdle, W. (1990), *Applied Nonparmetric Regression*, Cambridge University Press, Cambridge, England.

Jennen-Steinmetz, C. and Gasser, T. (1988), 'A unifying approach to nonparametric regression estimation', *Journal of the American Statistical Association* **83**(404), 1084–1089.

Kogure, A. (1987), 'Asymptotically optimal cells for a histogram', *Annals of Statistics* **15**, 1023–1030.

Li, Q. and Racine, J. S. (2007), *Nonparametric Econometrics: Theory and Practice*, Princeton University Press.

Loh, W.-Y. (2002), 'Regression trees with unbiased variable selection and interaction detection', *Statistica Sinica* **12**(2), 361–386.

Loh, W.-Y. (2011), 'Classification and regression trees', *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **1**(1), 14–23.

Mohlin, E. (2014*a*), A metric for measurable partitions. Mimeo.

Mohlin, E. (2014*b*), 'Optimal categorization', *Journal of Economic Theory* **152**, 356–381.

Morgan, J. M. and Sonquist, J. A. (1963), 'Problems in the analysis of survey data, and a proposal', *Journal of the American Statistical Association* **58**, 415–434.

Nadaraya, E. (1964), 'On estimating regression', *Theory of Probability and Its Applications* **9**(1), 141–142.

Priestley, M. B. and Chao, M. T. (1972), 'Non-parametric function fitting', *Journal of the Royal Statistical Society. Series B* **34**, 385–392.

Quinlan, J. R. (1992), Learning with continuous classes, *in* 'Proceedings of the 5th Australian joint Conference on Artificial Intelligence', Vol. 92, Singapore, pp. 343–348.

Ripley, B. D. (1996), *Pattern Recognition and Neural Networks*, Cambridge University Press, Cambridge.

Scott, D. W. (1992), *Multivariate Density Estimation*, John Wiley, New York.

Tukey, J. W. (1947), 'Non-parametric estimation ii. statistically equivalent blocks and tolerance regions–the continuous case', *The Annals of Mathematical Statistics* **18**(4), 529–539.

Watson, G. S. (1964), 'Smooth regression analysis', *Sankhyā: The Indian Journal of Statistics, Series A* pp. 359–372.

7 Appendix

7.1 Proofs: Preliminaries

Proof of Lemma 1.

$$\begin{aligned} ISE\left(\mathcal{C},t\right) &= \mathbb{E}\left[\left(Y|X-\bar{y}\left(X\right)\right)^2\right] \\ &= \sum_i p^i \mathbb{E}\left[\left(Y|X-\bar{y}^i\right)^2|X\in\mathcal{C}^i\right] \\ &= \sum_i p^i \mathbb{E}\left[\left(Y|X-\mu^i\right)^2+\left(\bar{y}^i-\mu^i\right)^2-2\left(Y|X-\mu^i\right)\left(\bar{y}^i-\mu^i\right)|X\in\mathcal{C}^i\right] \\ &= \sum_i p^i\left(Var\left(Y^i\right)+\left(\bar{y}^i-\mu^i\right)^2\right) \end{aligned}$$

■

Proof of Lemma 2. Suppose $x\in\mathcal{C}^i$. Start by noting

$$\begin{aligned} MSE\left(\mathcal{C},n\right)\left(x\right) &= \mathbb{E}\left[\left(Y\left(x\right)-\bar{Y}^i\right)^2\right] \\ &= \mathbb{E}\left[\left(\left(Y\left(x\right)-m\left(x\right)\right)^2+\left(m\left(x\right)-\bar{Y}^i\right)^2+2\left(Y\left(x\right)-m\left(x\right)\right)\left(m\left(x\right)-\bar{Y}^i\right)\right)\right] \\ &= \sigma^2\left(x\right)+\mathbb{E}\left[\left(m\left(x\right)-\bar{Y}^i\right)^2\right] \end{aligned}$$

Using $m\left(x\right)-\bar{Y}^i = m\left(x\right)-\mu^i-\left(\bar{Y}^i-\mu^i\right)$ we have

$$\mathbb{E}\left[\left(m\left(x\right)-\bar{Y}^i\right)^2\right] = Var\left(\bar{Y}^i\right)+\left(m\left(x\right)-\mu^i\right)^2.$$

Furthermore

$$Var\left(\bar{Y}^i\right) = \sum_{r=1}^{n}\Pr\left(N^i=r\right)\mathbb{E}\left[\left(\bar{Y}^i-\mu^i\right)^2|N^i=r\right]+\Pr\left(N^i=0\right)\mathbb{E}\left[\left(\bar{Y}_n-\mu^i\right)^2|N^i=0\right].$$

Note that if $r\geq 1$ then $E\left[\bar{Y}^i|N^i=r\right]=\mu^i$, so

$$\mathbb{E}\left[\left(\bar{Y}^i-\mu^i\right)^2|N^i=r\right] = Var\left(\bar{Y}^i|N^i=r\right) = \frac{1}{r}Var\left(Y^i\right).$$

It is evident that the number of objects in a cell, N^i, has a binomial distribution with parameters p^i and n. ■

Proof of Lemma 3. It is straightforward to verify that

$$\begin{aligned} Var\left(Y^i\right) &= \int_{x\in\mathcal{C}^i} \sigma^2\left(x\right) f_X\left(x|x\in\mathcal{C}^i\right) dx + \int_{x\in\mathcal{C}^i} \left(m\left(x\right)-\mu^i\right)^2 f_X\left(x|x\in\mathcal{C}^i\right) dx \qquad (16)\\ &= \mathbb{E}\left[\sigma^2\left(X\right)|X\in\mathcal{C}^i\right] + Var\left(m\left(X\right)|X\in\mathcal{C}^i\right). \end{aligned}$$

Using this we have

$$\begin{aligned} MISE\left(\mathcal{C},n\right) &= \sum_i p^i \int_{x\in\mathcal{C}^i} \left(\sigma^2\left(x\right) + Var\left(\bar{Y}^i\right) + \left(m\left(x\right)-\mu^i\right)^2\right) f_X\left(x|x\in\mathcal{C}^i\right) dx\\ &= \sum_i p^i \left(\mathbb{E}\left[\sigma^2\left(X\right)|X\in\mathcal{C}^i\right] + Var\left(\bar{Y}^i\right) + Var\left(m\left(X\right)|X\in\mathcal{C}^i\right)\right). \end{aligned}$$

■

7.2 Proofs: *MSE*- and *MISE* Optimality

Lemma 4 was proved in Mohlin (2014*b*).

7.2.1 Fixed design

Proof of Theorem 1. By lemma 4 an asymptotically optimal tree must satisfy $k\to\infty$ and $k/n\to 0$ as $n\to\infty$. This allows us to use lemma 2, to write, for $x\in\mathcal{C}^i$,

$$MSE\left(\mathcal{C},n\right)\left(x\right) = \sigma^2\left(x\right) + \underbrace{\left(m\left(x\right)-\mu^i\right)^2}_{\left(Bias\left[\bar{Y}(x)\right]\right)^2} + \underbrace{Var\left(Y^i\right)\frac{1}{np^i}}_{Var\left[\bar{Y}(x)\right]}. \qquad (17)$$

Also note

$$Var\left(Y^i\right) = \int_{s\in\mathcal{C}^i} \left[\sigma^2\left(s\right) + \left(m\left(s\right)-\mu^i\right)^2\right] f_X\left(s|s\in\mathcal{C}^i\right) ds. \qquad (18)$$

A Taylor approximation yields,

$$\begin{aligned} p^i &= \int_{a_d^i}^{b_d^i} ... \int_{a_1^i}^{b_1^i} \left[f_X\left(s\right)\right] ds_1...ds_d\\ &= \int_{a_d^i}^{b_d^i} ... \int_{a_1^i}^{b_1^i} \left[f_X\left(x\right) + D\left(f_X\left(x\right)\right)\left(s-x\right) + ...\right] ds_1...ds_d. \end{aligned}$$

Note that if $x_j \neq a_j^i + h_j/2$, then

$$\int_{a_j^i}^{b_j^i} (s_j - x_j)\, ds_j = \left(a_j^i + \frac{h_j}{2} - x_j\right) h_j^i = \left(\frac{h_j}{2} + O(h_j)\right) h_j = O\left(h_j^2\right),$$

and if $x_j = a_j^i + h_j/2$, then this is zero, so if there is at least one j such that $x_j \neq a_j^i + h_j/2$, then

$$\begin{aligned}\int_{a_d^i}^{b_d^i} ... \int_{a_1^i}^{b_1^i} \left[D\left(f_X(x)\right)(s-x)\right] ds_1...ds_d &= \int_{a_d^i}^{b_d^i} ... \int_{a_1^i}^{b_1^i} \left[\sum_j \frac{\partial f_X(x)}{\partial x_j}(s_j - x_j)\right] ds_1...ds_d \\ &= O\left(\times_{l=1}^d h_l\right) O(h_j),\end{aligned}$$

and if $x_j = a_j^i + h_j/2$, for all j, then this is zero. Moreover, it can be verified that

$$\int_{a_l^i}^{b_l^i} \frac{1}{2}(s-x) D^2\left(f_X(x)\right)(s-x)\, ds_l = \frac{1}{2}\sum_{i\neq l}\sum_{j\neq l} \frac{\partial^2 f_X(x)}{\partial x_j \partial x_i}(s_i - x_i)(s_j - x_j) h_l + O\left(h_l^3\right),$$

so

$$\begin{aligned}&\int_{a_d^i}^{b_d^i} ... \int_{a_1^i}^{b_1^i} \left[\frac{1}{2}(s-x) D^2\left(f_X(x)\right)(s-x)\right] ds_1...ds_d \\ &= \left(\int_{x_1-h_1/2}^{x_1+h_1/2} \frac{\partial^2 f_X(x)}{\partial^2 x_1}(s_1 - x_1)^2 ds_1\right) h_2 h_3 ... h_d + ... \\ &+ h_1 h_3 ... h_{n-1}\left(\int_{x_d-h_2/2}^{x_d+h_2/2} \frac{\partial^2 f_X(x)}{\partial^2 x_d}(s_d - x_d)^2 ds_d\right) \\ &= O\left(\times_{l=1}^d h_l \cdot \sum_j h_j^2\right).\end{aligned}$$

Thus if there is some j with $x_j \neq a_j^i + h_j/2$, then

$$p^i = \left(\times_{l=1}^d h_l\right)\left[f_X(x)\right] + O\left(\times_{l=1}^d h_l \cdot h_j\right),$$

and of not then

$$p^i = \left(\times_{l=1}^d h_l\right)\left[f_X(x)\right] + \left(\times_{l=1}^d h_l \cdot \sum_j h_j^2\right).$$

Consequently, we can assume

$$\frac{1}{np^i} = \frac{1}{f_X(x) \cdot n \cdot \left(\times_{l=1}^d h_l\right)} + O\left(\left(\times_{l=1}^d h_l \cdot h_j \cdot n\right)^{-1}\right). \tag{19}$$

Another Taylor approximation yields

$$\mu^i = \int_{s \in \mathcal{C}^i} \left[m(x) + D(m(x))(s-x) + \frac{1}{2}(s-x) D^2(m(x))(s-x) + ... \right] f_X\left(s | s \in \mathcal{C}^i\right) ds.$$

Note that

$$\int_{a_j^i}^{b_j^i} s_j f_X\left(s_j | s_j \in \mathcal{C}_j^i\right) ds_j = \mathbb{E}\left[s_j | s_j \in \mathcal{C}_j^i\right],$$

so

$$\int_{a_d^i}^{b_d^i} ... \int_{a_1^i}^{b_1^i} \left[\frac{\partial m(x)}{\partial x_j} (s_j - x_j) f_X\left(s | s \in \mathcal{C}^i\right) \right] ds_1 ... ds_d = \frac{\partial m(x)}{\partial x_j} \delta_j \left(\times_{l=1}^n h_l\right).$$

and

$$\begin{aligned}
&\int_{a_d^i}^{b_d^i} ... \int_{a_1^i}^{b_1^i} \left[D(m(x))(s-x) f_X\left(s | s \in \mathcal{C}^i\right) \right] ds_1 ... ds_d \\
&= \int_{a_d^i}^{b_d^i} ... \int_{a_1^i}^{b_1^i} \left[\sum_j \frac{\partial m(x)}{\partial x_j} (s_j - x_j) f_X\left(s | s \in \mathcal{C}^i\right) \right] ds_1 ... ds_d \\
&= \left(\times_{l=1}^n h_l\right) \sum_j \frac{\partial m(x)}{\partial x_j} \delta_j.
\end{aligned}$$

Furthermore, note that

$$\int_{a_d^i}^{b_d^i} ... \int_{a_1^i}^{b_1^i} \left[\frac{1}{2}(s-x) D^2(m(x))(s-x) f_X\left(s | s \in \mathcal{C}^i\right) \right] ds_1 ... ds_d = O\left(\left(\times_{l=1}^d h_l\right)^2\right).$$

Thus,

$$\mu^i = m(x) + \left(\times_{l=1}^d h_l\right) \sum_j \frac{\partial m(x)}{\partial x_j} \delta_j + O\left(\left(\times_{l=1}^d h_l\right)^2\right),$$

and consequently

$$\left(m(x) - \mu^i\right)^2 = \left(\left(\times_{l=1}^d h_l\right) \sum_j \frac{\partial m(x)}{\partial x_j} \delta_j \right)^2 + O\left(\left(\times_{l=1}^d h_l\right)^3\right). \tag{20}$$

Next use this to find

$$\begin{aligned}
&\int_{s\in\mathcal{C}^i}\left(m\left(s\right)-\mu^i\right)^2 f_X\left(s|s\in\mathcal{C}^i\right)ds \qquad (21)\\
&=\left(\times_{l=1}^d h_l\right)^2\int_{a_d^i}^{b_d^i}...\int_{a_1^i}^{b_1^i}\left(\sum_j\frac{\partial m\left(x\right)}{\partial x_j}\delta_j\right)^2 f_X\left(s|s\in\mathcal{C}^i\right)ds_1...ds_d\\
&+O\left(\left(\times_{l=1}^d h_l\right)^3\right)
\end{aligned}$$

Finally, a Taylor approximation yields

$$\int_{s\in\mathcal{C}^i}\sigma^2\left(s\right)f_X\left(s|s\in\mathcal{C}^i\right)ds=\sigma^2\left(x\right)+\sum_j\frac{\partial\sigma^2\left(x\right)}{\partial x_j}\left(\times_{l=1}^d h_l\right)\delta_j+O\left(\left(\times_{l=1}^d h_l\right)^3\right). \qquad (22)$$

Thus, using (20) and (21), equation (18) becomes,

$$Var\left(Y^i\right)=\sigma^2\left(x\right)+O\left(\times_{l=1}^d h_l\right).$$

Using (18)-(22) in (17) gives (7). This can be simplified to

$$MSE\left(\mathcal{C},n\right)\left(x\right)=\sigma^2\left(x\right)+h^2\left(\sum_j\frac{\partial m\left(x\right)}{\partial x_j}\delta_j\right)^2+\frac{\sigma^2\left(x\right)}{f_X\left(x\right)nh}. \qquad (23)$$

Maximise w.r.t. h. The first order condition is

$$\frac{\partial MSE\left(\mathcal{C},n\right)\left(x\right)}{\partial h}=2h\left(\sum_j\frac{\partial m\left(x\right)}{\partial x_j}\delta_j\right)^2-\frac{\sigma^2\left(x\right)}{f_X\left(x\right)}\frac{1}{nh^2}=0,$$

or (8), and the second order condition is satisfied,

$$\frac{\partial^2 MSE\left(\mathcal{C},n\right)\left(x\right)}{\partial h^2}=\frac{\sigma^2\left(x\right)}{nf_X\left(x\right)}\frac{1}{h^3}>0.$$

Evaluating (23) at (8) yields (9). ■

The proof of the claims made in remark 3 are omitted, but available upon request.

7.2.2 Random Design

The proof of theorem 2 relies on lemmas 6 and 7 below.

Lemma 6 *If $n\to\infty$, $k\to\infty$, and $k/n\to 0$, then $\sum_{r=1}^n \Pr\left(N^i=r\right)/r=O\left(1/np^i\right)$.*

Proof of Lemma 6. Note that $k/n \to 0$ implies $np^i \to \infty$. A Chernoff inequality for the binomial distribution (lemma 2.1 of Chung and Lu 2002) states that, for any real number λ,

$$\Pr\left(N \leq tp - \lambda\right) \leq e^{\frac{-\lambda^2}{2tp_i}},$$

so for $\lambda = \frac{1}{2}np^i$,

$$\Pr\left(N \leq \frac{1}{2}tp\right) \leq e^{\frac{-np^i}{8}}.$$

Let $r^* = \left\lfloor \frac{1}{2}np^i \right\rfloor$. Thus

$$\begin{aligned}
\sum_{r=1}^{n} \Pr\left(N^i = r\right) \frac{1}{r} &= \sum_{r=1}^{r^*-1} \Pr\left(N^i = r\right) \frac{1}{r} + \sum_{r=r^*}^{n} \Pr\left(N^i = r\right) \frac{1}{r} \\
&\leq \sum_{r=1}^{r^*-1} \Pr\left(N^i = r\right) + \sum_{r=r^*}^{n} \Pr\left(N^i = r\right) \frac{1}{r^*} \\
&\leq \Pr\left(N \leq r^* - 1\right) + \frac{1}{r^*} \\
&\leq e^{\frac{-np^i}{8}} + \frac{1}{\left\lfloor \frac{1}{2}np^i \right\rfloor} \\
&\leq e^{\frac{-np^i}{8}} + \frac{1}{\frac{1}{2}np^i - 1}.
\end{aligned}$$

Since $np^i/e^{np^i} \to 0$, as $np^i \to \infty$, we have

$$O\left(e^{\frac{-np^i}{8}} + \frac{1}{\frac{1}{2}np^i - 1}\right) = O\left(\frac{1}{\frac{1}{2}np^i - 1}\right) = O\left(\frac{1}{np^i}\right).$$

■

Lemma 7 *If $n \to \infty$, $k \to \infty$, and $k/n \to 0$, then, for all i,*

$$\Pr\left(N^i = 0\right) \mathbb{E}\left[\left(\bar{Y}_n - \mu^i\right)^2 | N^i = 0\right] = O\left(e^{-\frac{n}{k}}\right).$$

Proof of Lemma 7. Using $\mathbb{E}\left[\bar{Y}_n | N^i = 0\right] = \mu_{-i}$, we have

$$\begin{aligned}
\mathbb{E}\left[\left(\bar{Y}_n - \mu^i\right)^2 | N^i = 0\right] &= \mathbb{E}\left[\left(\bar{Y}_n - \mu_{-i}\right)^2 | N^i = 0\right] + \left(\mu_{-i} - \mu^i\right)^2 \\
&= Var\left(\bar{Y}_n | N^i = 0\right) + \left(\mu_{-i} - \mu^i\right)^2 \\
&= \frac{1}{n} Var\left(Y | X \notin \mathcal{C}^i\right) + \left(\mu_{-i} - \mu^i\right)^2. \qquad (24)
\end{aligned}$$

Next consider $\Pr\left(N^i=0\right)=\left(1-p^i\right)^n$. Let $p_{\max}=\max_i p^i$ and $p_{\min}=\min_i p^i$. Note that $p_{\min}>1-(k-1)\,p_{\max}$. Since $p_{\min}/\rho\geq p_{\max}$ we have $p_{\min}>1-(k-1)\,p_{\min}/\rho$, or equivalently $p_{\min}>\rho/\left(\rho+(k-1)\right)$. Hence

$$\left(1-p^i\right)^n\leq\left(1-p_{\min}\right)^n\leq\left(\left(1-\frac{\rho}{\rho+k-1}\right)^{(\rho+k-1)}\right)^{\frac{n}{\rho+k-1}}.$$

As $\rho+k-1\to\infty$ this approaches $e^{-\rho\frac{n}{\rho+k-1}}$. The desired result follows. ■

Proof of Theorem 2. By lemma 4 an asymptotically optimal tree must satisfy $k\to\infty$ and $k/n\to 0$ as $n\to\infty$. This allows us to use lemma 6 and lemma 7 together with lemma 2, to write, for $x\in\mathcal{C}^i$,

$$MSE\left(\mathcal{C},n\right)(x)=\sigma^2(x)+\underbrace{\left(m(x)-\mu^i\right)^2}_{Bias\left[\bar{Y}(x)\right]}+\underbrace{Var\left(Y^i\right)O\left(\frac{1}{p^in}\right)+O\left(e^{-\frac{n}{k}}\right)}_{Var\left[\bar{Y}(x)\right]}. \tag{25}$$

Using (18)-(22) (from the proof of theorem 1) in (25) gives

$$\begin{aligned}MSE\left(\mathcal{C},n\right)(x)=\sigma^2(x)+h^2\left(\sum_j\frac{\partial m(x)}{\partial x_j}\delta_j\right)^2+O\left(h^3\right)\\+\frac{\sigma^2(x)}{f_X(x)}O\left((n\cdot h)^{-1}\right)+O\left(n^{-1}\right)+O\left(e^{-\frac{n}{k}}\right).\end{aligned} \tag{26}$$

From above, we know that asymptotically $p^i=O(h)$, and $k^{-1}=O(h)$ so that $O\left(e^{-\frac{n}{k}}\right)=O\left(e^{-np^i}\right)$. Since $np^i\cdot e^{-np^i}\to 0$ this means that (26) can simplified to (10) or

$$MSE\left(\mathcal{C},n\right)(x)=\sigma^2(x)+h^2\left(\sum_j\frac{\partial m(x)}{\partial x_j}\delta_j\right)^2+\frac{\sigma^2(x)}{f_X(x)}O\left((nh)^{-1}\right).$$

Maximise w.r.t. h. The first order condition is

$$2h\left(\sum_j\frac{\partial m(x)}{\partial x_j}\delta_j\right)^2=\frac{\sigma^2(x)}{f_X(x)}O\left(n^{-1}h^{-2}\right),$$

and the second order condition is satisfied, so optimally (11). Using this in (10) yields (12). Having established that $h=O\left(n^{-2/3}\right)$ it can be shown that $e^{-\frac{n}{k}}$ vanishes faster than h^3 and n^{-1}. ■

7.2.3 Asymptotic Normality

Proof of Theorem 3. Since

$$\mathbb{E}\left[\bar{Y}(x)\right] - m(x) = Bias\left(\bar{Y}(x)\right) = h^2\left(\sum_j \frac{\partial m(x)}{\partial x_j}\delta_j\right)^2 + O\left(h^3\right),$$

we have, for $x \in \mathcal{C}^i$,

$$\begin{aligned}
&\sqrt{\frac{n^i}{Var\left(Y^i\right)}}\left[\bar{Y}(x) - m(x) - h^2\left(\sum_j \frac{\partial m(x)}{\partial x_j}\delta_j\right)^2\right] \\
&= \sqrt{\frac{n^i}{Var\left(Y^i\right)}}\left[\bar{Y}(x) - \mathbb{E}\left[\bar{Y}(x)\right]\right] + \sqrt{\frac{n^i}{Var\left(Y^i\right)}}\left[\mathbb{E}\left[\bar{Y}(x)\right] - m(x) - h^2\left(\sum_j \frac{\partial m(x)}{\partial x_j}\delta_j\right)^2\right] \\
&= \sqrt{\frac{n^i}{Var\left(Y^i\right)}}\left[\bar{Y}(x) - \mathbb{E}\left[\bar{Y}(x)\right]\right] + O\left(\sqrt{n^i}h^3\right) \\
&= \sqrt{\frac{n^i}{Var\left(Y^i\right)}}\sum_{Y^s \in t^i}\left[\frac{Y^s}{n^i} - \mathbb{E}\left[\frac{Y^s}{n^i}\right]\right] + o(1) \\
&= \sqrt{\frac{n^i}{Var\left(Y^i\right)}}\left[\sum_{Y^s \in t^i}\frac{Y^s}{n^i} - \mathbb{E}\left[\sum_{Y^s \in t^i}\frac{Y^s}{n^i}\right]\right] + o(1).
\end{aligned}$$

We now use the Liapunov Central Limit Theorem (e.g. Lemma A.5 in Li and Racine 2007). Define $Z_{n,s} \equiv Y^s/n^i$ and $S_n = \sum_{Y^s \in t^i} Z_{n,s}$ so that $\mathbb{E}[Z_{n,s}] = \mu^i/n^i$ and $Var(Z_{n,s}) = Var(Y^s)/(n^i)^2 = Var(Y^i)/(n^i)^2$, and $Var(S_n) = \sum_{Y^s \in t^i} Var(Z_{n,s}) = Var(Y^i)/n^i$. Since Y^s is bounded we have $\mathbb{E}|Z_{n,s}|^{2+q} < \infty$ for some $q > 0$. From (16) we see that

$$Var\left(Y^i\right) = \mathbb{E}\left[\sigma^2(X)\,|X \in \mathcal{C}^i\right] + o(1).$$

Finally note that

$$\lim_{n^i \to \infty}\sum_{Y^s \in t^i}\mathbb{E}\left|Z_{n,s} - \mathbb{E}[Z_{n,s}]\right|^{2+q} = 0,$$

for some $q > 0$. It follows that the Liapunov Central Limit Theorem implies

$$\sqrt{\frac{n^i}{Var\left(Y^i\right)}}\left[\sum_{Y^s \in t^i}\frac{Y^s}{n^i} - \mathbb{E}\left[\sum_{Y^s \in t^i}\frac{Y^s}{n^i}\right]\right] \xrightarrow{d} N(0,1),$$

or

$$\sqrt{n^i}\left[\sum_{Y^s\in t^i}\frac{Y^s}{n^i}-\mathbb{E}\left[\sum_{Y^s\in t^i}\frac{Y^s}{n^i}\right]\right]\xrightarrow{d}N\left(0,\sigma^2\left(x\right)\right).$$

■

7.3 Proofs: *ISE*-Optimality

Proof of Theorem 4. **(a)** Follows directly from the facts that

$$\mathbb{E}_{T\in\mathcal{T}(\mathcal{C},n)}\left[S^{i2}\right]=\mathbb{E}_{T\in\mathcal{T}(\mathcal{C},n)}\left[\frac{1}{n^i-1}\sum_{s\in T^i}\left(Y^s-\bar{Y}^i\right)^2\right]=Var\left(Y^i\right),$$

and

$$\mathbb{E}_{T\in\mathcal{T}(\mathcal{C},t)}\left[\left(\bar{Y}^i-\mu^i\right)^2\right]=\frac{1}{n^i}Var\left(Y^i\right).$$

(b) First we prove the result concerning $ISE\left(\mathcal{C},t\right)$. Consider a cell $\mathcal{C}_i\in\mathcal{C}$. Suppose that $n^i\geq 1$. It can be verified that

$$\left(\bar{Y}^i-\mu^i\right)^2=\left(\frac{1}{n^i}\sum_{s\in t^i}Y^s\right)^2+\mu^i-2\mu^i\frac{1}{n^i}\sum_{s\in t^i}Y^s.$$

Let $n^i\to\infty$. Since, for each cell i, $\{Y_s\}$ is an i.i.d. sequence with $\mathbb{E}\left[Y_s\right]=\mu^i$ we can use Kinchine's law of large numbers and Slutsky's lemma to conclude that $P\lim_{n^i\to\infty}\left(\bar{Y}^i-\mu^i\right)^2$ $=\mu^i+\mu^i-2\mu^i\mu^i=0$. In other words, for any $\varepsilon>0$ and $\delta\in(0,1)$, there is an $\bar{n}$ such that if $n^i>\bar{n}$ then $\Pr\left(\left(\bar{Y}^i-\mu^i\right)^2<\varepsilon\right)>\delta^{1/2}$. Moreover, for any $\bar{n}$ there is a n such that if $n>n$ then $\Pr\left(N^i>\bar{n}\right)>\delta^{1/2}$. This implies that, for any $\varepsilon>0$ and $\delta\in(0,1)$, there is a n such that if $n>n$ then $\Pr\left(\left(\bar{Y}^i-\mu^i\right)^2<\varepsilon\right)>\delta$. Thus, for any $\varepsilon>0$ and $\delta\in(0,1)$, there is a n such that if $n>n$ then

$$\Pr\left(\sum_{i=1}^{k}p^i\left(\bar{Y}^i-\mu^i\right)^2<\varepsilon\right)>\delta.$$

Since $ISE\left(\mathcal{C},t\right)\geq 0$, the desired result follows.

The proof for $SISE\left(\mathcal{C},t\right)$ is similar to the proof for $ISE\left(\mathcal{C},t\right)$, using the standard result $P\lim_{n^i\to\infty}s_i^2=Var\left(Y^i\right)$. ■

7.4 Proofs: Binary Outcomes

The proof of lemma 5 is the same as second half of the proof of lemma 2, and the proof of theorem 5 is essentially the same as the proof of theorem 7.