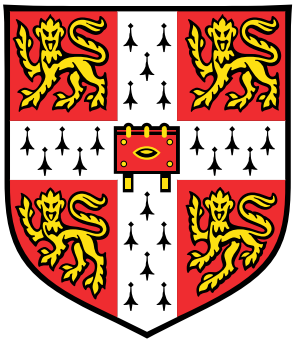
New Approaches to Modern Statistical Classification Problems



Timothy Ivor Cannings

Queens' College and Statistical Laboratory

University of Cambridge

August 2015

A thesis submitted for the degree of

Doctor of Philosophy

New Approaches to Modern Statistical Classification Problems

Timothy Ivor Cannings

Queens' College and Statistical Laboratory

University of Cambridge

A thesis submitted for the degree of

Doctor of Philosophy

This thesis concerns the development and mathematical analysis of statistical procedures for classification problems. In supervised classification, the practitioner is presented with the task of assigning an object to one of two or more classes, based on a number of labelled observations from each class. With modern technological advances, vast amounts of data can be collected routinely, which creates both new challenges and opportunities for statisticians. After introducing the topic and reviewing the existing literature in Chapter 1, we investigate two of the main issues to arise in recent times.

In Chapter 2 we introduce a very general method for high-dimensional classification, based on careful combination of the results of applying an arbitrary base classifier on random projections of the feature vectors into a lower-dimensional space. In one special case that we study in detail, the random projections are divided into non-overlapping blocks, and within each block we select the projection yielding the smallest estimate of the test error. Our random projection ensemble classifier then aggregates the results after applying the chosen projections, with a data-driven voting threshold to determine the final assignment. We derive bounds on the test error of a generic version of the ensemble as the number of projections increases. Moreover, under a low-dimensional boundary assumption, we show that the test error can be controlled by terms that do not depend on the original data dimension. The classifier is compared empirically with several other popular classifiers via an extensive simulation study, which reveals its excellent finite-sample performance.

Chapter 3 focuses on the k-nearest neighbour classifier. We first derive a new global asymptotic expansion for its excess risk, which elucidates conditions under which the dominant contribution to the risk comes from the locus of points at which each class label is equally likely to occur, as well as situations where the dominant contribution comes from the tails of the marginal distribution of the features. The results motivate an improvement to the k-nearest neighbour classifier in semi-supervised settings. Our proposal allows k to depend on an estimate of the marginal density of the features based on the unlabelled training data, using fewer neighbours when the estimated density at the test point is small. We show that the proposed semi-supervised classifier achieves a better balance in terms of the asymptotic local bias-variance trade-off. We also demonstrate the improvement in terms of finite-sample performance of the tail adaptive classifier over the standard classifier via a simulation study.

Preface

The main theme of this thesis – the development of new statistical procedures for the modern era – is, as is the case for many works in statistics nowadays, motivated by advances in technology. In particular, the facility to collect vast amounts of data routinely has created both new challenges and opportunities for statisticians. Primarily, can we make efficient use of all the data recorded?

At the beginning of the 21st century we entered the *digital age*. Analog formats, such as printed text and VHS tapes, were overtaken as the primary storage format. By 2007, 94% of data stored globally was in digital form, on PC hard disks and DVDs (Hilbert and López, 2011).

More recently, we have experienced what may be regarded as a media storm around *big data*, a broad term used to refer to any situation where one has more data than traditional techniques can cope with. As an example of the scale of the problems encountered, consider the *Square Kilometre Array* telescope[1], which is capable of recording more than an exabyte (10^{18} bytes) of raw data per day. Even processing and storing that amount of data is severely problematic!

The naive application of traditional statistical procedures in the context of big data may often lead to false conclusions. There are also ethical uncertainties, regarding privacy and legislation. Mayer-Schönberger and Cukier (2013) provide an accessible overview of many of the issues encountered. The big data era is likely still in its salad days. If we are to realise anywhere near the advertised potential, as statisticians, we must provide the pathway from (big) data to knowledge.

There are many people without whom this thesis would have most likely remained incomplete. First and foremost, it has been a privilege to work with Richard Samworth. His encouragement, guidance and patience have led to vast improvements in the contents of this work, and my development as a researcher. I would like to thank my friends and colleagues in the Statistical Laboratory, including John Aston, Tom Berrett, Julia Blackwell, Alexandra Carpentier, Nayia Constantinou, Yining Chen, Robin Evans, Arlene Kim, Danning Li, Will Matthews, Susan Pitts, Rajen Shah, John Shimmon, Tengyao Wang and Yi Yu, who have helped to provide a stimulating and

[1] `https://www.skatelescope.org`

friendly working environment over the past four years. Thank you also to all the friends who have taken an interest (or otherwise) in my work, especially Alan Sola, Peter Logg, Ed Barsley, Max Cooper, Tom Clarke, Jon Wilcox, Ellie Buckell, Thomas Whitcombe, Oliver Southwick, David Sykes, Thomas Hendicott, William Balfour, Damon Civin, Meline Joaris, Charlie Perrott and Sarah Gedye.

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University of similar institution.

Chapters 2 and 3 are joint work with Richard J. Samworth. A slightly shorter version of Chapter 2 has been submitted to the *Journal of the Royal Statistical Society, Series B*, and is available as Cannings and Samworth (2015a). We present here a more extensive simulation study. The corresponding `R` package `RPEnsemble` is available from `CRAN` (Cannings and Samworth, 2015b). We intend to submit Chapter 3 for publication soon.

Timothy Ivor Cannings

Cambridge, August 2015

Contents

Preface **v**

1 Introduction **1**

1.1 Thesis overview 4
1.2 Statistical setting 6
1.3 Traditional classification methods 8
1.3.1 Nonparametric methods 8
1.3.2 Parametric methods 9
1.3.3 Combinatorial methods 10
1.3.4 Bagging 11
1.4 Error estimation 11
1.5 Unbalanced class priors 12

2 Random projection ensemble classification **15**

2.1 Introduction 15
2.2 A generic random projection ensemble classifier 18
2.3 Choosing good random projections 21
2.4 Possible choices of the base classifier 23
2.4.1 Linear Discriminant Analysis 24
2.4.2 Quadratic Discriminant Analysis 26
2.4.3 The k-nearest neighbour classifier 27
2.5 Practical considerations 28
2.5.1 Choice of α 28
2.5.2 Choice of d 29
2.5.3 Choice of B_1 and B_2 32
2.6 Empirical analysis 32
2.6.1 Simulated examples 33
2.6.2 Real data examples 38
2.7 Discussion and extensions 39
2.8 Appendix 42

2.9 R code 57

3 Semi-supervised classification **65**

3.1 Introduction 65

3.2 Statistical setting 68

3.2.1 Preliminary result 69

3.3 Global risk of the k-nearest neighbour classifier 70

3.4 Tail adaptive classification 73

3.4.1 Oracle classifier 73

3.4.2 The semi-supervised nearest neighbour classifier 75

3.5 Empirical analysis 76

3.6 Appendix 79

3.6.1 Asymptotic expansions 81

3.6.2 Tail adaptive results 99

Chapter 1

Introduction

In a classification problem, the practitioner is presented with the task of assigning an object to one of two or more classes, based on a number of previous observations from each class. Many everyday decision problems fit into this framework. Traditional examples include, an email filter determining whether or not a message is spam, diagnoses of a disease based on the symptoms of a patient, or identifying a fraudulent financial transaction.

The *supervised* setting, where all the previous observations are labelled, was introduced by Fisher (1936). He developed Fisher's *Linear Discriminant Analysis* (LDA), and applied it to his *Iris* dataset, identifying the species of Iris plants based on measurements of the sepals and petals; see Figure 1.1. The LDA method, and the related *Quadratic Discriminant Analysis* (QDA), is still widely used today in forming the basis of many modern classification techniques.

Moving on from Fisher's work, the 1950s and '60s was a revolutionary period for statistics. Rosenblatt (1956) and Parzen (1962) introduced a rigorous framework for the analysis of nonparametric methods. In fact, Fix and Hodges (1951) (later republished as Fix and Hodges, 1989) proposed a nonparametric method for classification. This would later become the well-known k-nearest neighbour classifier (Cover and Hart, 1967). Furthermore, there was the development of Vapnik–Chervonenkis (VC) Theory (Vapnik and Chervonenkis, 1971), which facilitates distribution-free analysis of learning techniques. Then, with the production of the microprocessor and subsequent widespread use of the electronic computer in the 1970s, there was an influx of data, generating unprecedented demand for new, faster statistical procedures. Classification was one of the central problems arising.

With more recent technological advances, complex data structures are commonplace. The scope of applications extends far beyond the everyday problems mentioned above. Observations may be high-resolution 3D fMRI images, full gene-expression sequences, or accelerometer recordings from wearable technology. The resulting problems

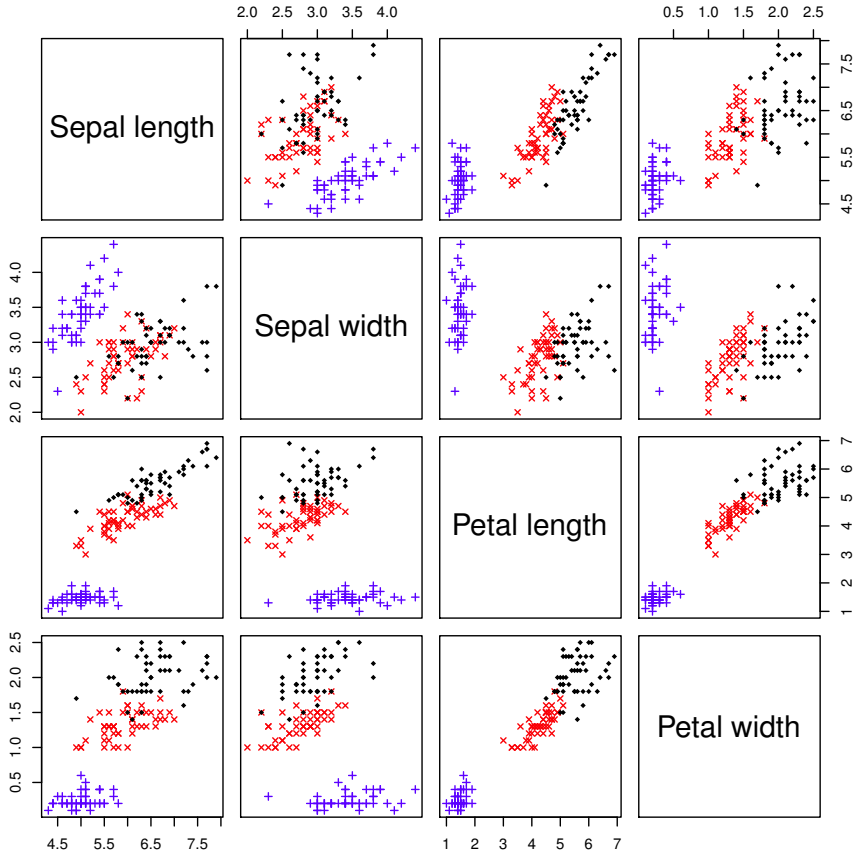


Figure 1.1: *A scatter plot of Fisher's Iris dataset. The three different species, namely versicolor, setosa, and virginica, are represented by the red, blue and black points, respectively.*

for the statistician are numerous. There may simply be too much data to directly apply a traditional method or the data may only be accessible for a short time; one may need to combine data from many different sources, and there may be a large number nuisance (or noise) variables.

A comparison of Figures 1.1 and 1.2 further elucidates some of these issues. Note first that the plot of the Iris data is very informative, the classes are nicely clustered and the distribution of each class is well approximated by a Normal distribution, suggesting that Fisher's LDA should work well (cf. Section 1.3.2). This is not the case for the latter figure, which presents the Cardiac Arrhythmia[1] data (available from the UCI machine learning repository). The task here is to identify and, if present, characterise a patient's irregular heartbeat, based on a number of electrocardiogram (ECG) recordings. The plot in Figure 1.2 presents just 5 of the 279 variables and 6 of the 16 classes recorded in the dataset.

The topic of classification receives interest not only from the mainstream statistical

[1] `https://archive.ics.uci.edu/ml/datasets/Arrhythmia`

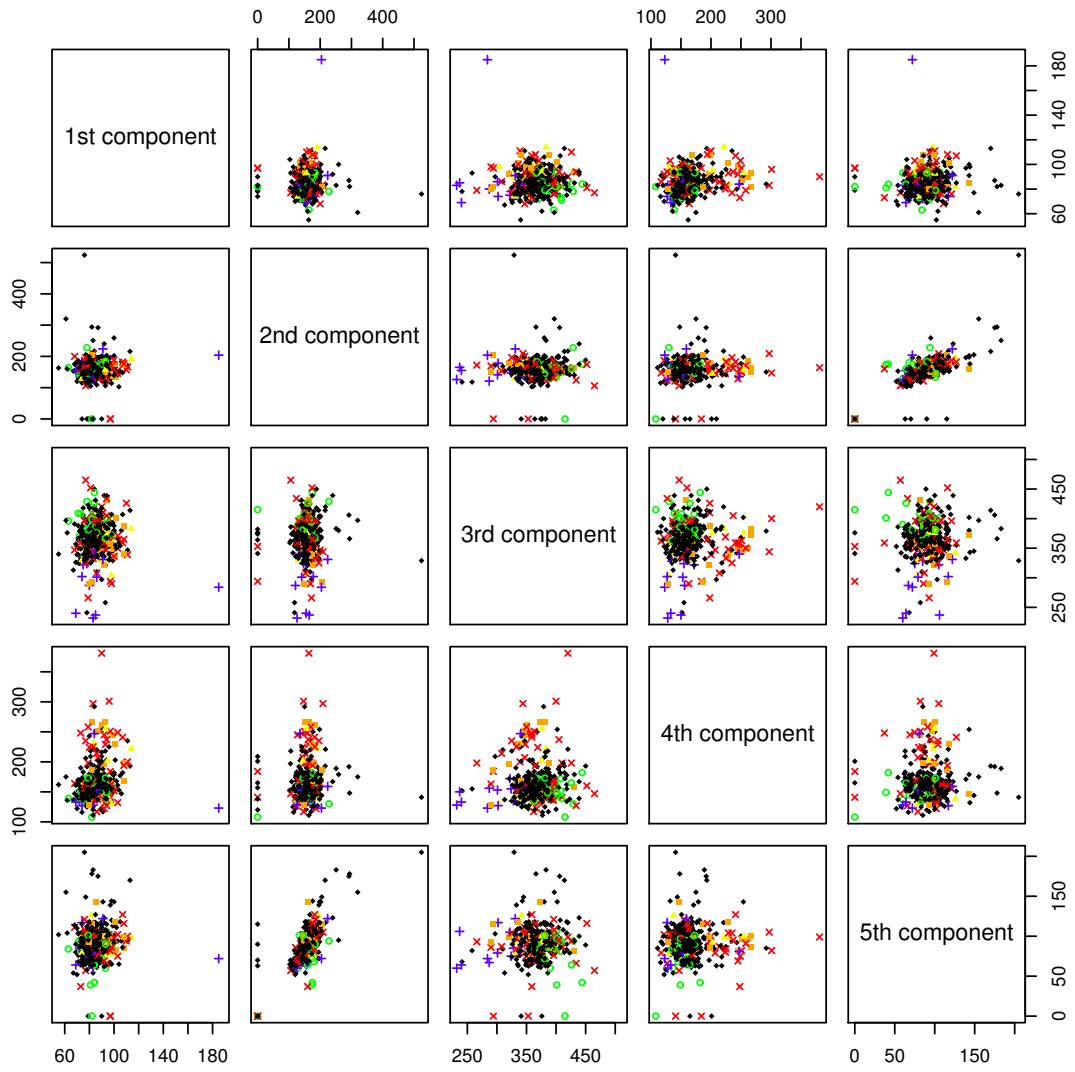


Figure 1.2: *A scatter plot of a subset of the Cardiac Arrhythmia dataset. The colours represent six different types of Cardiac Arrhythmia. The five components shown are different ECG recordings.*

community, but also computer scientists and engineers. In fact, *The British Classification Society*[2], which celebrated its 50 anniversary in 2014, has members from diverse fields such as anthropology, astronomy, biology, forensic science and psychology. With such interdisciplinary approaches, the literature has advanced rapidly over the last decade. A search for the phrase "statistical classification" on Google Scholar returns 3560 results from the last year alone[3]. Other particularly active classification societies of note include the *Gesellschaft für Klassifikation e.V.*[4] and *The Section of Classifica-*

[2] `http://www.brclasssoc.org.uk`
[3] Search performed on 31 July 2015.
[4] `http://www.gfkl.org/en/`

tion and Data Analysis[5] of the Polish Statistical Association. Moreover, the *International Federation of Classification Societies*[6] - currently consisting of 16 international member societies - host a biennial conference, publish a newsletter and support two journals dedicated to the subject, namely the *Journal of Classification* and *Advances in Data Analysis and Classification*.

The practitioner now has perhaps thousands of classification methods at their disposal. Hastie, Tibshirani and Friedman (2009) provide a good overview of popular statistical learning procedures, including many for classification. There are also many books focussing on classification, notably Hand (1981), Devroye, Györfi and Lugosi (1996), Gordon (1999) and Duda, Hart and Stork (2000). The survey paper by Boucheron, Bousquet and Lugosi (2005) provides a substantial review of the relevant literature up to 2005. A more up to date survey can be found in Fan, Fan and Wu (2010).

1.1 Thesis overview

The two main chapters in this thesis focus on different aspects of modern classification problems. In Chapter 2, we introduce a new method for high-dimensional data, where the dimension p of the feature vectors may be comparable to or even greater than the number of training data points, n. In such settings, classical methods such as those mentioned in Section 1.3 tend to perform poorly (Bickel and Levina, 2004), and may even be intractable; for example, this is the case for LDA, where the problems are caused by the fact that the sample covariance matrix is not invertible when $p \geq n$.

Many methods proposed to overcome such problems assume that the optimal decision boundary between the classes is linear, e.g. Friedman (1989) and Hastie et al. (1995). Another common approach assumes that only a small subset of features are relevant for classification. Examples of works that impose such a sparsity condition include Fan and Fan (2008), where it is also assumed that the features are independent, as well as Tibshirani et al. (2003) and Guo, Hastie and Tibshirani (2007), where soft thresholding is used to obtain a sparse boundary.

More recently, Witten and Tibshirani (2011) and Fan, Feng and Tong (2012) both solve an optimisation problem similar to Fisher's linear discriminant, with the addition of an ℓ_1 penalty term to encourage sparsity. Other works in this area include Cai and Liu (2011); Clemmensen et al. (2011); Fan et al. (2015) and Hao et al. (2015). An attractive property of many of these methods is their interpretability; often a list of the most important features will be returned along with the class assignments. This

[5] `http://www.us.szc.pl/main.php/skad_ang/`

[6] `http://ifcs.boku.ac.at`

information may be very useful to the practitioner, for example, it may enable a doctor to effectively target their treatment of a particular disease.

In Chapter 3, we propose a *semi-supervised* k-nearest neighbour classifier, where the number of neighbours considered varies with the location of the test point. More precisely, we first estimate the marginal density of the features using a large unlabelled training data set, then let k depend on this estimate at the test point, using fewer neighbours when the density is small.

An increasing number of modern classification problems are of so-called semi-supervised form, where only some of the observations are labelled, while others (often the large majority) are unlabelled. Such problems typically occur because acquiring the label associated with a large set of observations may be particularly time-consuming or expensive, such as determining whether there is oil at a potential drilling site, or whether historical bank transactions are fraudulent or legitimate.

An overview of semi-supervised learning techniques can be found in Chapelle et al. (2006). Arguably their most successful application in classification problems is in *density-based metric* methods. For instance, when the features take values in some lower-dimensional manifold, one can use the unlabelled training data to first learn the manifold and use this structure to improve on the naive method that works directly in the larger ambient space. As another example, when using the k-nearest neighbour classifier, one may use the unlabelled samples to determine a more suitable distance metric (Friedman, 1994; Bijral et al., 2012). Azizyan et al. (2013) provide a rigorous statistical framework for the analysis of semi-supervised methods of this type; see also Liang et al. (2007).

Another common condition used in semi-supervised classification problems is the *cluster assumption*. This assumption states that the *high-density* regions of feature space can be written as a countable disjoint union of measurable sets or clusters, on which the Bayes classifier (defined in (1.1) below) is constant. For instance, Rigollet (2007) first uses the unlabelled data to estimate the clusters, then the labelled training data to provide information for the classification for each set. The unlabelled data, in contrast to the proposal here, do not help with classification in the low density regions.

Before presenting the main chapters, we outline the statistical setting used in this thesis, provide a brief review of some traditional classification methods, and highlight two under-developed aspects of classification problems.

1.2 Statistical setting

The setting used throughout this thesis is as follows: For $n, m \in \mathbb{N}$, we have independent random pairs $(X, Y), (X_1, Y_1), \ldots, (X_{n+m}, Y_{n+m})$ taking values in[7] $\mathbb{R}^p \times \{1, \ldots, K\}$, each with joint distribution P. We will focus on the binary classification problem ($K = 2$), and briefly mention the extension to the general K case where relevant. In supervised problems, we observe X and the *training* data pairs $\mathcal{T}_n := \{(X_1, Y_1), \ldots (X_n, Y_n)\}$, and would like to predict the class label Y. In the semi-supervised setting, we further observe the feature vectors $X_{n+1}, \ldots, X_{n+m}$.

Alternatives to this setting, which are not considered here, include the *Poisson* model (see Hall et al., 2008), which assumes the pairs are generated from a marked Poisson process. In this case, the sample size is not fixed, reflecting a situation where observations arise sequentially, consider, for example, a bank looking to detect fraud in financial transactions. The *discriminant analysis* model assumes that the class sizes are fixed in advance, so there is no randomness in the class labels (see, for example, Mammen and Tsybakov, 1999). Such a setting would arise if the training data pairs were generated by experimental design. For instance, a doctor may choose a fixed number of healthy and unhealthy patients to form the training set.

In our setting, the joint distribution P can be characterised in two ways. We may first generate the class label from the marginal distribution of Y. That is Y takes the value 1 with probability π_1 and the 2 with probability $\pi_2 := 1 - \pi_1$. Then, conditional on $Y = r$, X has distribution P_r, for $r = 1, 2$. Alternatively, we may first generate X from its marginal distribution $P_X := \pi_1 P_1 + \pi_2 P_2$; then, conditional on $X = x$, the probability that $Y = 1$ is given by the regression function $\eta(x) := \mathbb{P}(Y = 1 | X = x)$. To formally define $\eta : \mathbb{R}^p \to [0, 1]$, for a Borel set $B \subseteq \mathbb{R}^p \times \{1, 2\}$, write

$$B = \{B \cap (\mathbb{R}^p \times \{1\})\} \cup \{B \cap (\mathbb{R}^p \times \{2\})\} =: (B_1 \times \{1\}) \cup (B_2 \times \{2\}),$$

then

$$\mathbb{P}\{(X, Y) \in B\} = \int_{B_1} \eta(x)\, dP_X(x) + \int_{B_2} \{1 - \eta(x)\}\, dP_X(x).$$

Alternatively,

$$\mathbb{P}\{(X, Y) \in B\} = \pi_1 \int_{B_1} dP_1(x) + \pi_2 \int_{B_2} dP_2(x).$$

Thus, specifying either π_1, P_1 and P_2, or P_X and η, will fix the joint distribution P.

A *classifier* is a measurable function $C : \mathbb{R}^p \to \{1, 2\}$, with the interpretation that C assigns x to the class $C(x)$. We let $\mathcal{C}_p$ denote the set of all such classifiers. The central task in this thesis is to find classifiers with small misclassification rate, or *risk*,

[7] In fact, in Chapter 3 we denote the dimension of feature vectors by d, following an informal convention that p and d refer to high and low dimensions, respectively.

over a prescribed set $\mathcal{R} \subseteq \mathbb{R}^p$, given by $R_{\mathcal{R}}(C) := \mathbb{P}\{C(X) \neq Y, X \in \mathcal{R}\}$. Here, the set $\mathcal{R}$ may be the whole of $\mathbb{R}^d$, in which case the subscript $\mathcal{R}$ will be dropped, or a particular subset of $\mathbb{R}^d$ of interest.

The *Bayes* classifier

$$C^{\text{Bayes}}(x) = \begin{cases} 1 & \text{if } \eta(x) \geq 1/2; \\ 2 & \text{otherwise,} \end{cases} \tag{1.1}$$

(e.g. Devroye, Györfi and Lugosi, 1996, p. 10) minimises the risk over any set $\mathcal{R}$. Its risk is

$$\begin{aligned} R_{\mathcal{R}}(C^{\text{Bayes}}) &= \int_{\mathcal{R}} \eta(x)\mathbb{1}_{\{\eta(X)<1/2\}} + \{1-\eta(x)\}\mathbb{1}_{\{\eta(X)\geq 1/2\}}\, dP_X(x) \\ &= \mathbb{E}[\min\{\eta(X), 1-\eta(X)\}\mathbb{1}_{\{X\in\mathcal{R}\}}]. \end{aligned}$$

The set of points $\mathcal{S} := \{x \in \mathbb{R}^p : \eta(x) = 1/2\}$ is commonly referred to as the *Bayes decision boundary*. Of course, in practice the regression function is unknown, so the the Bayes classifier cannot be used directly.

A classifier $\hat{C}_n$, based on the n labelled training data points, is a measurable function from $(\mathbb{R}^p \times \{1,2\})^n$ to $\mathcal{C}_p$. A number of statistical questions naturally arise regarding the non-negative excess risk, or *regret*, given by $R(\hat{C}_n) - R(C^{\text{Bayes}})$:

- Can we bound the difference between $\lim_{n\to\infty} R(\hat{C}_n)$ and $R(C^{\text{Bayes}})$?
- Is $\hat{C}_n$ *consistent*, i.e. does the excess risk converge to 0 as $n \to \infty$?
- How fast does the excess risk of a consistent classifier converge as n (and possibly p) $\to \infty$?
- Are there finite sample bounds on the excess risk?

An elegant result relating to the first question is for the 1-nearest neighbour (1nn) classifier, which assigns the point x to the class of its closest (according to some norm on $\mathbb{R}^p$) point in the training set. It can be shown that the asymptotic error of the 1nn classifier is $\mathbb{E}[2\eta(X)\{1-\eta(X)\}]$. Cover and Hart (1967) provide the following bounds

$$R(C^{\text{Bayes}}) \leq \mathbb{E}[2\eta(X)\{1-\eta(X)\}] \leq 2R(C^{\text{Bayes}})\{1-R(C^{\text{Bayes}})\} \leq 2R(C^{\text{Bayes}});$$

see also Devroye et al. (1996, p. 22). The 1nn classifier, then, potentially provides a lot of information about the Bayes risk, especially if it is small.

1.3 Traditional classification methods

In this section, we outline and discuss many classification methods and some of the theory developed for them.

1.3.1 Nonparametric methods

A large subclass of techniques are known as *plug-in* classifiers. The idea here is to first estimate the regression function nonparametrically, and then use the estimate in place of η in the Bayes classifier. Many methods fit into this framework, for example the k-nearest neighbour classifier, histogram classifiers and kernel classifiers.

Formally, let $\hat{\eta}_n$ denote the regression estimate, and define the plug-in classifier

$$\hat{C}_n(x) = \begin{cases} 1 & \text{if } \hat{\eta}_n(x) \geq 1/2; \\ 2 & \text{otherwise.} \end{cases} \tag{1.2}$$

In this case, the excess risk satisfies (Devroye et al., 1996, Theorem 2.2)

$$R(\hat{C}_n) - R(C^{\text{Bayes}}) = \mathbb{E}\{|2\eta(X) - 1||\mathbb{1}_{\{\hat{\eta}_n(X)<1/2\}} - \mathbb{1}_{\{\eta(X)<1/2\}}|\} \leq 2\mathbb{E}|\hat{\eta}_n(X) - \eta(X)|.$$

It follows that, if $\hat{\eta}_n$ is a consistent estimate of η, then the plug-in classifier based on $\hat{\eta}_n$ is also consistent. In fact, classification is an easier problem than regression, in the sense that the rate of convergence of the misclassification error will be faster than the L_2-error of the regression estimate (Devroye et al., 1996, Chapter 6.7).

To obtain precise rates of convergence, we must restrict the class of joint distributions of the pair (X, Y) considered. Two assumptions are commonly made. The first is a complexity or *smoothness* assumption, for example that the function η belongs to a Hölder class with smoothness parameter β_0. The second is the so-called *margin assumption*, which states that there exists $C_0 > 0$ and $\alpha_0 \geq 0$, such that

$$\mathbb{P}\{0 < |\eta(X) - 1/2| \leq t\} = C_0 t^{\alpha_0},$$

for all $t > 0$ (Mammen and Tsybakov, 1999; Tsybakov, 2004; Audibert and Tsybakov, 2007).

These two assumptions characterise the difficulty of a classification problem. Intuitively, since we need only determine whether or not $\eta > 1/2$, when η is far from $1/2$, classification is easy. That is, the problem is easier for larger α_0. Moreover, if η is smooth, then estimating it will be easier. Audibert and Tsybakov (2007) derive the rate of convergence of a general plug-in classifier. In particular, it is shown, under a further regularity assumption on the distribution P_X, that a local polynomial estimator

of η can achieve a $O(n^{-\beta_0(1+\alpha_0)/(2\beta_0+p)})$ rate of convergence. They show further that this is the minimax rate for this class of distributions.

Arguably the most popular plug-in classifier is the k-nearest neighbour ($k$nn) classifier, which estimates η at $x \in \mathbb{R}^p$ by a majority vote over the classes of the k nearest neighbours to x. This method has the attractive property of being *universally consistent* (Stone, 1977). That is, so long as $k := k_n$ diverges with n, but $k_n/n \to 0$, then the $k$nn classifier is consistent for any distribution P. Furthermore, Hall et al. (2008) showed that the $k$nn classifier, with an optimal choice of k, can achieve the minimax rate above when $\alpha_0 = 1$ and $\beta_0 = 2$. Samworth (2012a) studies the more general *weighted nearest neighbour* classifier, where decreasing weights are given to more distant neighbours. An optimal weighting scheme is derived, under which the resulting classifier is guaranteed to outperform its vanilla $k$nn counterpart asymptotically.

As one can see presented above, the rate of convergence of the excess risk of the $k$nn classifier (and for general nonparametric plug-in classifiers) is slow when p is large. This is often referred to as suffering from the *curse of dimensionality*.

1.3.2 Parametric methods

Suppose we have a parametric model for each of the conditional class distributions. Then one can simply estimate the parameters (via maximum likelihood, or otherwise) and if the model is a good approximation the resulting classifier will perform well.

Fisher's original proposal followed this framework. Recall that in the case where $X|Y = r \sim N_p(\mu_r, \Sigma)$, we have

$$\mathrm{sgn}\{\eta(x) - 1/2\} = \mathrm{sgn}\Big\{\log\frac{\pi_1}{\pi_2} + \Big(x - \frac{\mu_1+\mu_2}{2}\Big)^T\Sigma^{-1}(\mu_1 - \mu_2)\Big\}.$$

In LDA, π_r, μ_r and Σ are estimated by their sample versions, using a pooled estimate of Σ. We then define

$$\hat{C}_n^{\mathrm{LDA}}(x) := \begin{cases} 1 & \text{if } \log\frac{\hat{\pi}_1}{\hat{\pi}_2} + \big(x - \frac{\hat{\mu}_1+\hat{\mu}_2}{2}\big)^T\hat{\Omega}(\hat{\mu}_1 - \hat{\mu}_2) \geq 0; \\ 2 & \text{otherwise,} \end{cases} \tag{1.3}$$

where $\hat{\Omega} := \hat{\Sigma}^{-1}$. When $p > n$ the estimate of Σ is singular and the LDA method cannot be applied directly.

Under this model, when $\pi_1 = \pi_2 = 1/2$ we have that $R(C^{\mathrm{Bayes}}) = \Phi\big(-\|\Sigma^{-1/2}(\mu_1 - \mu_2)\|\big)$, where Φ denotes the standard Normal distribution function. The fact the Bayes risk can be written in closed form facilitates more straightforward analysis.

1.3.3 Combinatorial methods

Combinatorial methods do not rely on a statistical model. Instead, the idea is to divide the feature space into regions which classify the training data well.

One simple and intuitive idea is *empirical risk minimisation*. Suppose we have a set $\mathcal{C}$ of potential classifiers. To pick a classifier from $\mathcal{C}$, we choose the one with the smallest empirical error over the training set,

$$\hat{C}_n^* := \operatorname*{argmin}_{C \in \mathcal{C}} \Big\{ \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{C(X_i) \neq Y_i\}} \Big\}.$$

This idea was developed significantly in a series of papers by Vapnik and Chervonenkis (1971, 1974a,b), which led in part to what is now known as Vapnik–Chervonenkis (VC) Theory; see also Devroye et al. (1996, Chapters 12-14) and Vapnik (2000).

Let $\mathcal{L}_n(C) := \mathbb{P}\{C(X) \neq Y | \mathcal{T}_n\}$ denote the *test error* of the classifier C. By writing

$$\mathcal{L}_n(\hat{C}_n^*) - R(C^{\text{Bayes}}) = \big\{ \mathcal{L}_n(\hat{C}_n^*) - \inf_{C \in \mathcal{C}} \mathcal{L}_n(C) \big\} + \big\{ \inf_{C \in \mathcal{C}} \mathcal{L}_n(C) - R(C^{\text{Bayes}}) \big\},$$

we see there is a trade-off in the choice of the set $\mathcal{C}$. If the class is large enough, then the second term will be small, but this may come at the cost of increasing the first term. More precisely, Vapnik and Chervonenkis provide the remarkable, distribution-free result (see Devroye et al., 1996, Theorem 12.6) that

$$\mathbb{P}\big\{ \mathcal{L}_n(\hat{C}_n^*) - \inf_{C \in \mathcal{C}} \mathcal{L}_n(C) > \epsilon \big\} \leq 8 S(\mathcal{C}, n) \exp(-n\epsilon^2/128).$$

Here, $S(\mathcal{C}, n)$ is the nth shatter coefficient of the class $\mathcal{C}$, which does not depend on the distribution P.

Suppose $\mathcal{C}$ is chosen to be the set of all linear classifiers on $\mathbb{R}^p$. In this case, it can be shown that $S(\mathcal{C}, n) = n^{p+1}$ (see Devroye et al., 1996, Chapter 13.3). As with plug-in classifiers, this method is not well suited to high-dimensional problems.

Another popular technique, motivated by the work of Vapnik and Chervonenkis in the 1960s on *optimal separating hyperplanes* (see Vapnik and Chervonenkis, 1974c), is the *support vector machine* (SVM) (Cortes and Vapnik, 1995). The SVM is constructed by finding an optimal separating hyperplane after mapping the feature vectors into a high-dimensional space via the so-called *kernel trick*. The method is easy to implement and can be very general, depending on the kernel chosen. The reader is referred to Vapnik (2000, Chapter 5) for an in-depth review of this topic.

1.3.4 Bagging

There are also techniques aimed at improving the performance of an existing method. One of the first ideas of this type was bootstrap aggregating, or *bagging* (Breiman, 1996), in which test point is classified after many bootstrap resamples of the training data, then the final assignment is made via a majority vote. Bagging was shown to be particularly effective in one setting by Hall and Samworth (2005): a bagged version of the typically inconsistent 1-nearest neighbour classifier is universally consistent. A more general rule of thumb states bagging a weak classifier can lead to vast improvements in performance; however, bagging a strong classifier is often futile.

Breiman's early work on bagging led to the development of the extremely popular Random Forest classifier (Breiman, 2001). This method combines bootstrap resampling with random feature selection, and classifies the test point using a tree (Breiman et al., 1984) grown on the chosen features. As in bagging, the individual trees are then combined via a majority vote. Whilst the method enjoys very good empirical performance, its theoretical properties are not well understood. One paper of note on this topic, however, is Biau et al. (2008), where it is shown that a variant of the original Random Forest proposal is consistent.

Another related idea is *boosting*, in which the importance, or weight, of each training data pair is updated at each stage of the ensemble. A particularly successful proposal is *Adaptive Boosting* (AdaBoost) by Freund and Schapire (1997). The idea here is to incrementally update the algorithm, giving an increased weight to the training points that were incorrectly classified during the previous step. The final ensemble then also weights the individual learners based on their error. As with the Random Forest method, it's not yet fully understood why boosting works so well. Bartlett and Traskin (2007) study the consistency properties of AdaBoost. A comprehensive review of boosting techniques can be found in Freund and Schapire (2012); see also Schapire (2013).

1.4 Error estimation

An important practical aspect of classification problems is error estimation. In particular, we would like to estimate the *test error*, $\mathcal{L}_n(\hat{C}_n) := \mathbb{P}\{\hat{C}_n(X) \neq Y | \mathcal{T}_n\}$. Of course, the distribution of the pair (X, Y) is unknown, and therefore so is $\mathcal{L}_n$. A simple way to estimate the error is to use a test set: consider a situation in which we also observe the pairs $(X_{n+1}, Y_{n+1}), \ldots, (X_{n+m}, Y_{n+m})$. Then a natural, unbiased error estimate is

$$\hat{L}_{n,m}(\hat{C}_n) := \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}_{\{\hat{C}_n(X_{n+i}) \neq Y_{n+i}\}}.$$

This is often used to estimate the error in simulation studies, where one can generate as many test pairs as desired. The estimate is very good for large m; more precisely, by Hoeffding's inequality (Devroye et al., 1996, p. 122), we have that

$$\mathbb{P}\big\{|\mathcal{L}_n(\hat{C}_n) - \hat{L}_{n,m}(\hat{C}_n)| \geq \epsilon \;\big|\; \mathcal{T}_n\big\} \leq 2e^{-2m\epsilon^2}.$$

In practical problems we rarely have a labelled test set. In some situations, it's possible that one may have enough training data to split the sample into a smaller training set and a test set. If this isn't the case, then more sophisticated methods are available. The simplest is the empirical error, or resubstitution, estimator $\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{\hat{C}_n(X_i)\neq Y_i\}}$. The problem here is the tendency to overfit to the training sample; for instance, the estimate is always zero for the 1-nearest neighbour classifier.

A method less susceptible to overfitting is the leave-one-out estimator, given by $\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{\hat{C}_{n,-i}(X_i)\neq Y_i\}}$, where $\hat{C}_{n,-i}$ denotes the classifier trained without the ith data pair. Corresponding results to the Hoeffding bound above can also be found for these estimators; see Devroye et al. (1996), Chapters 23 and 24.

Error estimates are also often used in order to select a classifier from a subset of possible choices, usually to choose a tuning parameter. A widely used method here is N-fold cross validation. The training sample is split into N approximately equal sized blocks at random, then the number of errors made on each block is counted when the classifier is trained with the remaining data after removing the current block. (Note that when $N = n$ we have the leave-one-out estimator.) This process is repeated for each choice of the parameter, and the one yielding the smallest estimate is retained.

A remaining problem is that of error uncertainty quantification. We would like to have a confidence interval for the error of a particular classifier, so that the practitioner, a doctor for example, can make an informed decision regarding a diagnoses. Laber and Murphy (2011) propose a method to calculate a confidence interval for the error of a linear empirical risk minimiser (c.f. Section 1.3.3), which adapts to the non-smoothness of the test error.

1.5 Unbalanced class priors

A major problem, perhaps somewhat neglected by the mainstream classification community, is the situation where one class is far more prevalent than the other. Primarily, there will be a lack of data from the minority class. One proposal is to repeatedly subsample the majority class, so that the class sizes are balanced. However, in extreme cases, the minority class may be so small that this method becomes unstable.

A further issue in this setting, is that the risk (as defined in Section 1.2) may not be a good measure of performance; the 'optimal' method may simply assign every

observation to the majority class. Some proposals attempt to minimise the area under the receiver operator characteristics (ROC) curve, that is the false positive-true positive rate (see, for example, Xue and Hall (2015)). However, such a criteria will to be too optimistic in some settings; we may do better by targeting a trade-off specific to the problem at hand.

Consider a scenario in which a doctor is diagnosing a rare disease. As a preliminary step, the doctor tests patients with the purpose of giving the all clear (class 1), or sending a patient for further (more intensive, expensive) testing, since they potentially have the disease (class 2). Now, the risk of a classifier C has the following expansion:

$$R(C) = \mathbb{P}\{C(X) \neq Y\} = \pi_1 \mathbb{P}\{C(X) = 2|Y = 1\} + \pi_2 \mathbb{P}\{C(X) = 1|Y = 2\}.$$

Since the disease is rare, we are expecting $\pi_1 \gg \pi_2$, so the risk is giving a much higher weight to sending a healthy patient for further testing than giving an unhealthy patient the all clear. In practice, the trade-off would likely be the other way round.

This poses a question to the statistician. What is a suitable measure of performance in this setting? It will likely depend on the problem at hand. There is desire, therefore, for flexible methods that can be easily adapted to target the specific criteria.

Chapter 2

Random projection ensemble classification

2.1 Introduction

In this chapter, we introduce a very general method for high-dimensional classification, based on careful combination of the results of applying an arbitrary base classifier to random projections of the feature vectors into a lower-dimensional space. In one special case that we study in detail, the random projections are divided into non-overlapping blocks, and within each block we select the projection yielding the smallest estimate of the test error. Our random projection ensemble classifier then aggregates the results of applying the base classifier on the selected projections, with a data-driven voting threshold to determine the final assignment.

The use of random projections in high-dimensional statistical problems is motivated by the celebrated Johnson–Lindenstrauss Lemma (e.g. Dasgupta and Gupta, 2002). This lemma states that, given $x_1, \ldots, x_n \in \mathbb{R}^p$, $\epsilon \in (0,1)$ and $d > \frac{8 \log n}{\epsilon^2}$, there exists a linear map $f : \mathbb{R}^p \to \mathbb{R}^d$ such that

$$(1-\epsilon)\|x_i - x_j\|^2 \leq \|f(x_i) - f(x_j)\|^2 \leq (1+\epsilon)\|x_i - x_j\|^2,$$

for all $i, j = 1, \ldots, n$. In fact, the function f that nearly preserves the pairwise distances can be found in randomised polynomial time using random projections distributed according to Haar measure as described in Section 2.3 below. It is interesting to note that the lower bound on d in the Johnson–Lindenstrauss Lemma does not depend on p. As a result, random projections have been used successfully as a computational time saver: when p is large compared to $\log n$, one may project the data at random into a lower-dimensional space and run the statistical procedure on the projected data, potentially making great computational savings, while achieving comparable or even

improved statistical performance. As one example of the above strategy, Durrant and Kaban (2013) obtained Vapnik–Chervonenkis type bounds on the generalisation error of a linear classifier trained on a single random projection of the data. See also Dasgupta (1999), Ailon and Chazelle (2006) and McWilliams et al. (2014) for other instances.

Other works have sought to reap the benefits of aggregating over many random projections. For instance, Marzetta, Tucci and Simon (2011) considered estimating a $p \times p$ population inverse covariance matrix using $B^{-1} \sum_{b=1}^{B} A_b^T (A_b \hat{\Sigma} A_b^T)^{-1} A_b$, where $\hat{\Sigma}$ denotes the sample covariance matrix and $A_1, \ldots, A_B$ are random projections from $\mathbb{R}^p$ to $\mathbb{R}^d$. Lopes, Jacob and Wainwright (2011) used this estimate when testing for a difference between two Gaussian population means in high dimensions, while Durrant and Kaban (2014) applied the same technique in Fisher's linear discriminant for a high-dimensional classification problem.

The main motivation here extends beyond the Johnson–Lindenstrauss Lemma. Suppose that X_1 and X_2 are random vectors taking values in $\mathbb{R}^p$. It follows that, by considering characteristic functions, if AX_1 and AX_2 are identically distributed for all projections A from $\mathbb{R}^p$ to $\mathbb{R}^d$, for a given $d \leq p$, then X_1 and X_2 are identically distributed. Put another way, if the distributions of X_1 and X_2 are different, then, for each $d \leq p$, there exists a projection $A_0 : \mathbb{R}^p \to \mathbb{R}^d$ such that the distributions of $A_0 X_1$ and $A_0 X_2$ are different.

Our proposed methodology for high-dimensional classification has some similarities with the techniques described above, in the sense that we consider many random projections of the data, but is also closely related to *bagging* (Breiman, 1996) and the *Random Forest* classifier (Breiman, 2001), since the ultimate assignment of each test point is made by aggregation and a vote. Bagging has proved to be an effective tool for improving unstable classifiers; indeed, a bagged version of the (generally inconsistent) 1-nearest neighbour classifier is universally consistent as long as the resample size is carefully chosen; see Hall and Samworth (2005). More generally, bagging has been shown to be particularly effective in high-dimensional problems such as variable selection (Meinshausen and Bühlmann, 2010; Shah and Samworth, 2013). Another related approach to ours is Blaser and Fryzlewicz (2015), who consider ensembles of random rotations, as opposed to projections.

One of the basic but fundamental observations that underpins our proposal is the fact that aggregating the classifications of all random projections is not sensible, since most of these projections will typically destroy the class structure in the data; see the top row of Figure 2.1. For this reason, we advocate partitioning the projections into non-overlapping blocks, and within each block we retain only the projection yielding the smallest estimate of the test error. The attraction of this strategy is illustrated in the bottom row of Figure 2.1, where we see a much clearer partition of the classes.

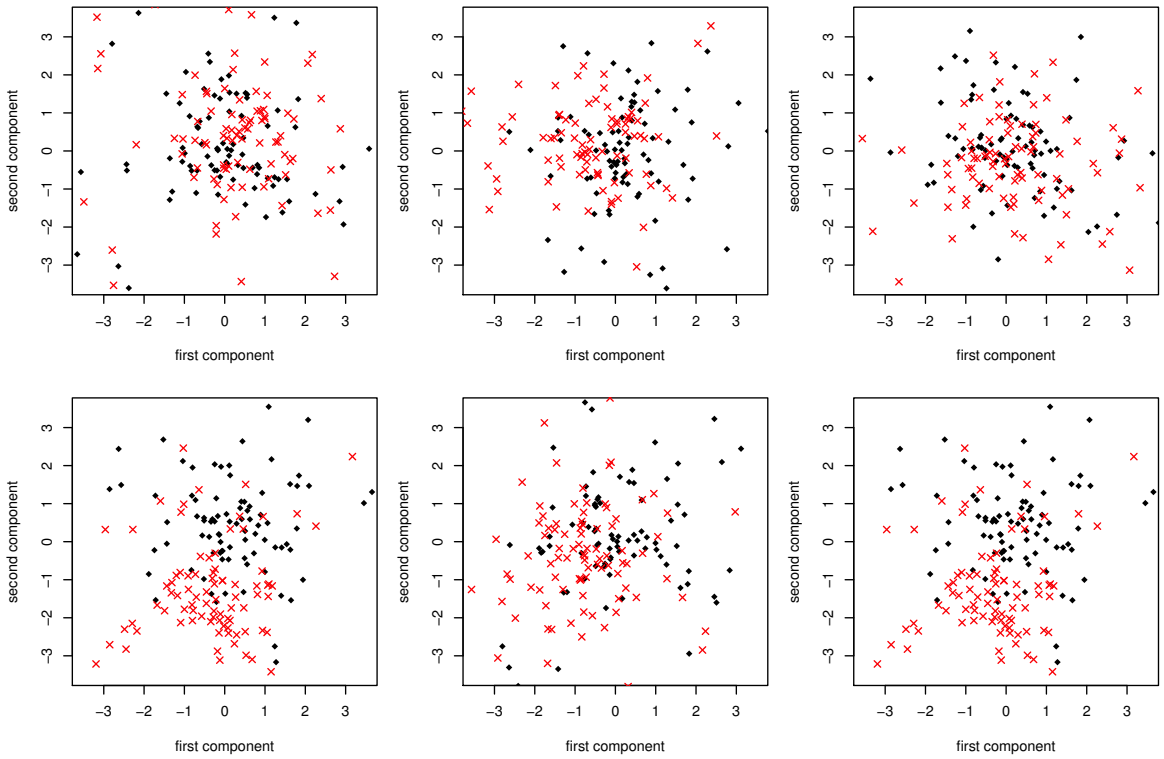


Figure 2.1: *Different two-dimensional projections of a sample of size* $n = 200$ *from Model 2 in Section 2.6.1 with* $p = 50$ *dimensions and prior probability* $\pi_1 = 1/2$. *Top row: three projections drawn from Haar measure; bottom row: the projections with smallest estimate of test error out of 100 Haar projections with LDA (left), QDA (middle) and k-nearest neighbours (right).*

Another key feature of our proposal is the realisation that a simple majority vote of the classifications based on the retained projections can be highly suboptimal; instead, we argue that the voting threshold should be chosen in a data-driven fashion in an attempt to minimise the test error of the infinite-simulation version of our random projection ensemble classifier. In fact, this estimate of the optimal threshold turns out to be remarkably effective in practice; see Section 2.5.1 for further details. We emphasise that our methodology can be used in conjunction with any base classifier, though we particularly have in mind classifiers designed for use in low-dimensional settings. The random projection ensemble classifier can therefore be regarded as a general technique for either extending the applicability of an existing classifier to high dimensions, or improving its performance.

Our theoretical results are divided into three parts. In the first, we consider a generic base classifier and a generic method for generating the random projections into $\mathbb{R}^d$ and quantify the difference between the test error of the random projection ensemble classifier and its infinite-simulation counterpart as the number of projections increases. We then consider selecting random projections from non-overlapping blocks by initially

drawing them according to Haar measure, and then within each block retaining the projection that minimises an estimate of the test error. Under a condition implied by the widely-used sufficient dimension reduction assumption (Li, 1991; Cook, 1998; Lee et al., 2013), we can then control the difference between the test error of the random projection classifier and the Bayes risk as a function of terms that depend on the performance of the base classifier based on projected data and our method for estimating the test error, as well as terms that become negligible as the number of projections increases. The final part of our theory gives risk bounds on the first two of these terms for specific choices of base classifier, namely Fisher's linear discriminant and the k-nearest neighbour classifier. The key point here is that these bounds only depend on d, the sample size n and the number of projections, and not on the original data dimension p.

The remainder of this chapter is organised as follows. Our methodology and general theory are developed in Sections 2.2 and 2.3. Specific choices of base classifier are discussed in Section 2.4, while Section 2.5 is devoted to a consideration of the practical issues of the choice of voting threshold, projected dimension and the number of projections used. In Section 2.6 we present results from an extensive empirical analysis on both simulated and real data where we compare the performance of the random projection ensemble classifier with several popular techniques for high-dimensional classification. The outcomes are extremely encouraging, and suggest that the random projection ensemble classifier has excellent finite-sample performance in a variety of different high-dimensional classification settings. We conclude with a discussion of various extensions and open problems. All proofs are deferred to the Appendix.

Finally in this section, we introduce the following general notation used throughout this chapter. For a sufficiently smooth real-valued function g defined on a neighbourhood of $t \in \mathbb{R}$, let $\dot{g}(t)$ and $\ddot{g}(t)$ denote its first and second derivatives at t, and let $\lfloor t \rfloor$ and $[\![t]\!] := t - \lfloor t \rfloor$ denote the integer and fractional part of t respectively.

2.2 A generic random projection ensemble classifier

We start by recalling our setting from Section 1.2 in the Introduction and defining the relevant notation. Suppose that the pair (X, Y) takes values in $\mathbb{R}^p \times \{1, 2\}$, with joint distribution P, characterised by $\pi_1 := \mathbb{P}(Y = 1)$, and P_r, the conditional distribution of $X|Y = r$, for $r = 1, 2$. For convenience, we let $\pi_2 := \mathbb{P}(Y = 2) = 1 - \pi_1$. In the alternative characterisation of P, we let P_X denote the marginal distribution of X and write $\eta(x) := \mathbb{P}(Y = 1|X = x)$ for the regression function. Recall that a *classifier* on $\mathbb{R}^p$ is a Borel measurable function $C : \mathbb{R}^p \to \{1, 2\}$, with the interpretation that we assign a point $x \in \mathbb{R}^p$ to class $C(x)$. We let $\mathcal{C}_p$ denote the set of all such classifiers.

The misclassification rate, or *risk*, of a classifier C is $R(C) := \mathbb{P}\{C(X) \neq Y\}$, and is minimised by the *Bayes* classifier

$$C^{\mathrm{Bayes}}(x) = \begin{cases} 1 & \text{if } \eta(x) \geq 1/2; \\ 2 & \text{otherwise.} \end{cases}$$

Of course, we cannot use the Bayes classifier in practice, since η is unknown. Nevertheless, we have access to a sample of training data that we can use to construct an approximation to the Bayes classifier. Throughout this section and Section 2.3, it is convenient to consider the training sample $\mathcal{T}_n := \{(x_1, y_1), \ldots, (x_n, y_n)\}$ to be fixed points in $\mathbb{R}^p \times \{1, 2\}$. Our methodology will be applied to a base classifier $\hat{C}_n = \hat{C}_{n,\mathcal{T}_{n,d}}$, which we assume can be constructed from an arbitrary training sample $\mathcal{T}_{n,d}$ of size n in $\mathbb{R}^d \times \{1, 2\}$; thus $\hat{C}_n$ is a measurable function from $(\mathbb{R}^d \times \{1, 2\})^n$ to $\mathcal{C}_d$.

Now assume that $d \leq p$. We say a matrix $A \in \mathbb{R}^{d\times p}$ is a *projection* if $AA^T = I_{d\times d}$, the d-dimensional identity matrix. Let $\mathcal{A} = \mathcal{A}_{d\times p} := \{A \in \mathbb{R}^{d\times p} : AA^T = I_{d\times d}\}$ be the set of all such matrices. Given a projection $A \in \mathcal{A}$, define projected data $z_i^A := Ax_i$ and $y_i^A := y_i$ for $i = 1, \ldots, n$, and let $\mathcal{T}_n^A := \{(z_1^A, y_1^A), \ldots, (z_n^A, y_n^A)\}$. The projected data base classifier corresponding to $\hat{C}_n$ is $\hat{C}_n^A : (\mathbb{R}^d \times \{1, 2\})^n \to \mathcal{C}_p$, given by

$$\hat{C}_n^A(x) = \hat{C}_{n,\mathcal{T}_n^A}^A(x) := \hat{C}_{n,\mathcal{T}_n^A}(Ax).$$

Note that although $\hat{C}_n^A$ is a classifier on $\mathbb{R}^p$, the value of $\hat{C}_n^A(x)$ only depends on x through its d-dimensional projection Ax.

We now define a generic ensemble classifier based on random projections. For $B_1 \in \mathbb{N}$, let $A_1, A_2, \ldots, A_{B_1}$ denote independent and identically distributed projections in $\mathcal{A}$, independent of (X, Y). The distribution on $\mathcal{A}$ is left unspecified at this stage, and in fact our proposed method ultimately involves choosing this distribution depending on $\mathcal{T}_n$. Now set

$$\hat{\nu}_n^{B_1}(x) := \frac{1}{B_1} \sum_{b_1=1}^{B_1} \mathbb{1}_{\{\hat{C}_n^{A_{b_1}}(x)=1\}}. \tag{2.1}$$

For $\alpha \in (0, 1)$, the *random projection ensemble* classifier is defined to be

$$\hat{C}_n^{\mathrm{RP}}(x) := \begin{cases} 1 & \text{if } \hat{\nu}_n^{B_1}(x) \geq \alpha; \\ 2 & \text{otherwise.} \end{cases} \tag{2.2}$$

We emphasise again here the additional flexibility afforded by not pre-specifying the voting threshold α to be $1/2$. Our analysis of the random projection ensemble classifier will require some further definitions. Let

$$\hat{\mu}_n(x) := \mathbb{E}\{\hat{\nu}_n^{B_1}(x)\} = \mathbb{P}\{\hat{C}_n^{A_1}(x) = 1\},$$

where the randomness here comes from the random projections. Let $G_{n,1}$ and $G_{n,2}$ denote the distribution functions of $\hat{\mu}_n(X)|\{Y=1\}$ and $\hat{\mu}_n(X)|\{Y=2\}$, respectively.

We will make use of the following assumption:

(A.1) $G_{n,1}$ and $G_{n,2}$ are twice differentiable at α.

The first derivatives of $G_{n,1}$ and $G_{n,2}$, when they exist, are denoted as $g_{n,1}$ and $g_{n,2}$ respectively; under **(A.1)**, these derivatives are well-defined in a neighbourhood of α. Our first main result below gives an asymptotic expansion for the test error $\mathcal{L}(\hat{C}_n^{\mathrm{RP}}) := \mathbb{P}\{\hat{C}_n^{\mathrm{RP}}(X) \neq Y\}$ of our generic random projection ensemble classifier as the number of projections increases. In particular, we show that this test error can be well approximated by the test error of the infinite-simulation random projection classifier

$$\hat{C}_n^{\mathrm{RP}^*}(x) := \begin{cases} 1 & \text{if } \hat{\mu}_n(x) \geq \alpha; \\ 2 & \text{otherwise.} \end{cases}$$

This infinite-simulation classifier turns out to be easier to analyse in subsequent results. Note that under **(A.1)**,

$$\mathcal{L}(\hat{C}_n^{\mathrm{RP}^*}) := \mathbb{P}\{\hat{C}_n^{\mathrm{RP}^*}(X) \neq Y\} = \pi_1 G_{n,1}(\alpha) + \pi_2\{1 - G_{n,2}(\alpha)\}. \tag{2.3}$$

Theorem 2.1. *Assume **(A.1)**. Then*

$$\mathcal{L}(\hat{C}_n^{\mathrm{RP}}) - \mathcal{L}(\hat{C}_n^{\mathrm{RP}^*}) = \frac{\gamma_n(\alpha)}{B_1} + o\Big(\frac{1}{B_1}\Big)$$

as $B_1 \to \infty$, where

$$\gamma_n(\alpha) := (1 - \alpha - [\![B_1\alpha]\!])\{\pi_1 g_{n,1}(\alpha) - \pi_2 g_{n,2}(\alpha)\} + \frac{\alpha(1-\alpha)}{2}\{\pi_1 \dot{g}_{n,1}(\alpha) - \pi_2 \dot{g}_{n,2}(\alpha)\}.$$

Lopes (2013) provides a similar conclusion for a majority vote over a general exchangeable sequence of base classifiers, satisfied by procedures such as bagging (cf. Section 1.3.4). The conditions of his result are different to those imposed here. The proof of Theorem 2.1 in the Appendix is lengthy, and involves a one-term Edgeworth approximation to the distribution function of a standardised Binomial random variable. One of the technical challenges is to show that the error in this approximation holds uniformly in the binomial proportion.

Define the test error of $\hat{C}_n^A$ by[1]

$$\mathcal{L}_n^A := \int_{\mathbb{R}^p \times \{1,2\}} \mathbb{1}_{\{\hat{C}_n^A(x) \neq y\}} \, dP(x,y).$$

[1]We define $\mathcal{L}_n^A$ through an integral rather than defining $\mathcal{L}_n^A := \mathbb{P}\{\hat{C}_n^A(X) \neq Y\}$ to make it clear that when A is a random projection, it should be conditioned on when computing $\mathcal{L}_n^A$.

Our next result controls the test excess risk, i.e. the difference between the test error and the Bayes risk, of the infinite-simulation random projection classifier in terms of the expected test excess risk of the classifier based on a single random projection. An attractive feature of this result is its generality: no assumptions are placed on the configuration of the training data $\mathcal{T}_n$, the distribution P of the test point (X, Y) or on the distribution of the individual projections.

Theorem 2.2. *We have*

$$\mathcal{L}(\hat{C}_n^{\mathrm{RP}^*}) - R(C^{\mathrm{Bayes}}) \leq \frac{1}{\min(\alpha, 1-\alpha)} \{\mathbb{E}(\mathcal{L}_n^{A_1}) - R(C^{\mathrm{Bayes}})\}.$$

2.3 Choosing good random projections

In this section, we study a special case of the generic random projection ensemble classifier introduced in Section 2.2, where we propose a screening method for choosing the random projections. Let $\hat{L}_n^A$ be an estimator of $\mathcal{L}_n^A$, based on the projected data $\{(z_1^A, y_1^A), \ldots, (z_n^A, y_n^A)\}$, that takes values in the set $\{0, 1/n, \ldots, 1\}$. Examples of such estimators include resubstitution and leave-one-out estimates; we discuss these choices in greater detail in Section 2.4. For $B_1, B_2 \in \mathbb{N}$, let $\{A_{b_1,b_2} : b_1 = 1, \ldots, B_1, b_2 = 1, \ldots, B_2\}$ denote independent projections, independent of (X, Y), distributed according to Haar measure on $\mathcal{A}$. One way to simulate from Haar measure on the set $\mathcal{A}$ is to first generate a matrix $R \in \mathbb{R}^{d \times p}$, where each entry is drawn independently from a standard normal distribution, and then take A^T to be the matrix of left singular vectors in the singular value decomposition of R^T. For $b_1 = 1, \ldots, B_1$, let

$$b_2^*(b_1) := \operatorname*{sargmin}_{b_2 \in \{1,\ldots,B_2\}} \hat{L}_n^{A_{b_1,b_2}}, \tag{2.4}$$

where sargmin denotes the smallest index where the minimum is attained in the case of a tie. We now set $A_{b_1} := A_{b_1, b_2^*(b_1)}$, and consider the random projection ensemble classifier from Section 2.2 constructed using the independent projections $A_1, \ldots, A_{B_1}$.

Let

$$\hat{L}_n^* := \min_{A \in \mathcal{A}} \hat{L}_n^A$$

denote the optimal test error estimate over all projections. The minimum is attained here, since $\hat{L}_n^A$ takes only finitely many values. For $j = 0, 1, \ldots, \lfloor n(1 - \hat{L}_n^*) \rfloor$, let

$$\beta_n(j) := \mathbb{P}\big(\hat{L}_n^A \leq \hat{L}_n^* + j/n\big),$$

where A is distributed according to Haar measure on $\mathcal{A}$. We assume the following:

(A.2) There exist $\beta_0 \in (0,1)$ and $\beta, \rho > 0$ such that

$$\beta_n(j) \geq \beta_0 + \frac{\beta j^\rho}{n^\rho}$$

for $j \in \big\{0, 1, \ldots, \big\lfloor n\big(\frac{\log^2 B_2}{\beta B_2}\big)^{1/\rho}\big\rfloor + 1\big\}$.

Condition **(A.2)** asks for a certain growth rate of the distribution function of $\hat{L}_n^A$ close to its minimum value $\hat{L}_n^*$; observe that the strength of the condition decreases as B_2 increases. Under this condition, the following result is a starting point for controlling the expected test excess risk of the classifier based on a single projection chosen according to the scheme described above.

Proposition 2.3. *Assume **(A.2)**. Then*

$$\mathbb{E}(\mathcal{L}_n^{A_1}) - R(C^{\mathrm{Bayes}}) \leq \hat{L}_n^* - R(C^{\mathrm{Bayes}}) + \epsilon_n + (1-\beta_0)^{B_2}\bigg\{\frac{1}{n} + \frac{(1-\beta_0)^{1/\rho}\Gamma(1+1/\rho)}{B_2^{1/\rho}\beta^{1/\rho}} + \exp\Big(-\frac{\log^2 B_2}{1-\beta_0}\Big)\bigg\},$$

where $\epsilon_n = \epsilon_n^{(B_2)} := \mathbb{E}(\mathcal{L}_n^{A_1} - \hat{L}_n^{A_1})$.

The form of the bound in Proposition 2.3 motivates us to seek to control $\hat{L}_n^* - R(C^{\mathrm{Bayes}})$ in terms of the test excess risk of a classifier based on the projected data, in the hope that we will be able to show this does not depend on p. To this end, define the regression function on $\mathbb{R}^d$ induced by the projection $A \in \mathcal{A}$ to be $\eta^A(z) := \mathbb{P}(Y = 1|AX = z)$. The corresponding induced Bayes classifier, which is the optimal classifier knowing only the distribution of (AX, Y), is given by

$$C^{A-\mathrm{Bayes}}(z) := \begin{cases} 1 & \text{if } \eta^A(z) \geq 1/2; \\ 2 & \text{otherwise.} \end{cases}$$

Its risk is

$$R^{A-\mathrm{Bayes}} := \int_{\mathbb{R}^p \times \{1,2\}} \mathbb{1}_{\{C^{A-\mathrm{Bayes}}(Ax) \neq y\}}\, dP(x,y).$$

In order to ensure that $\hat{L}_n^*$ will be close to the Bayes risk, we will invoke an additional assumption on the form of the Bayes classifier:

(A.3) There exists a projection $A^* \in \mathcal{A}$ such that

$$P_X(\{x \in \mathbb{R}^p : \eta(x) \geq 1/2\} \triangle \{x \in \mathbb{R}^p : \eta^{A^*}(A^*x) \geq 1/2\}) = 0,$$

where $B \triangle C := (B \cap C^c) \cup (B^c \cap C)$ denotes the symmetric difference of two sets B and C.

Condition **(A.3)** requires that the set of points $x \in \mathbb{R}^p$ assigned by the Bayes classifier to class 1 can be expressed as a function of a d-dimensional projection of x. Note that if the Bayes decision boundary is a hyperplane, then **(A.3)** holds with $d = 1$. Moreover, **(A.3)** holds if the problem is *sparse*; for example, suppose that only the first d of the p features are relevant for classification, then A^* can be taken to be the first d Euclidean basis vectors. Proposition 2.4 below shows that, in fact, **(A.3)** holds under the sufficient dimension reduction condition, which states that Y is independent of X given A^*X; see Cook (1998) for many statistical settings where such an assumption is natural.

Proposition 2.4. *If Y is independent of X given A^*X, then **(A.3)** holds.*

Finally, then, we are in a position to control the test excess risk of our random projection ensemble classifier in terms of the test excess risk of a classifier based on d-dimensional data, as well as terms that reflect our ability to estimate the test error of classifiers based on projected data and terms that depend on B_1 and B_2.

Theorem 2.5. *Assume **(A.1)**, **(A.2)** and **(A.3)**. Then*

$$\mathcal{L}(\hat{C}_n^{\mathrm{RP}}) - R(C^{\mathrm{Bayes}}) \leq \frac{\mathcal{L}_n^{A^*} - R^{A^*-\mathrm{Bayes}}}{\min(\alpha, 1-\alpha)} + \frac{\epsilon_n - \epsilon_n^{A^*}}{\min(\alpha, 1-\alpha)} + \frac{\gamma_n(\alpha)}{B_1}\{1 + o(1)\} + \frac{(1-\beta_0)^{B_2}}{\min(\alpha, 1-\alpha)}\biggl\{\frac{1}{n} + \frac{(1-\beta_0)^{1/\rho}\Gamma(1+1/\rho)}{B_2^{1/\rho}\beta^{1/\rho}} + \exp\Bigl(-\frac{\log^2 B_2}{1-\beta_0}\Bigr)\biggr\}$$

as $B_1 \to \infty$, where $\gamma_n(\alpha)$ is defined in Theorem 2.1, ϵ_n is defined in Proposition 2.3 and $\epsilon_n^{A^*} := \mathcal{L}_n^{A^*} - \hat{L}_n^{A^*}$.

Regarding the bound in Theorem 2.5 as a sum of four terms, we see that the last two of these can be seen as the price we have to pay for the fact that we do not have access to an infinite sample of random projections. These terms can be made negligible by choosing B_1 and B_2 to be sufficiently large, but it should be noted that ϵ_n may increase with B_2. This is a reflection of the fact that minimising an estimate of test error may lead to overfitting. The behaviour of this term, together with that of $\mathcal{L}_n^{A^*} - R^{A^*-\mathrm{Bayes}}$ and $\epsilon_n^{A^*}$, depends on the choice of base classifier, but in the next section below we describe settings where these terms can be bounded by expressions that do not depend on p.

2.4 Possible choices of the base classifier

In this section, we change our previous perspective and regard the training data $\mathcal{T}_n := \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ as independent random pairs with distribution P, so our earlier

statements are interpreted conditionally on $\mathcal{T}_n$. We consider particular choices of base classifier, and study the first two terms in the bound in Theorem 2.5.

2.4.1 Linear Discriminant Analysis

Linear Discriminant Analysis (LDA), introduced by Fisher (1936), is arguably the simplest classification technique. Recall that in the special case where $X|Y = r \sim N_p(\mu_r, \Sigma)$, we have

$$\mathrm{sgn}\{\eta(x) - 1/2\} = \mathrm{sgn}\biggl\{\log\frac{\pi_1}{\pi_2} + \Bigl(x - \frac{\mu_1+\mu_2}{2}\Bigr)^T\Sigma^{-1}(\mu_1-\mu_2)\biggr\},$$

so **(A.3)** holds with $d = 1$ and $A^* = \frac{(\mu_1-\mu_2)^T\Sigma^{-1}}{\|\Sigma^{-1}(\mu_1-\mu_2)\|}$, a $1\times p$ matrix. In LDA, π_r, μ_r and Σ are estimated by their sample versions, using a pooled estimate of Σ. Although LDA cannot be applied directly when $p \geq n$ since the sample covariance matrix is singular, we can still use it as the base classifier for a random projection ensemble, provided that $d < n$. Indeed, noting that for any $A \in \mathcal{A}$, we have $AX|Y = r \sim N_d(\mu_r^A, \Sigma^A)$, where $\mu_r^A := A\mu_r$ and $\Sigma^A := A\Sigma A^T$, we can define

$$\hat{C}_n^A(x) = \hat{C}_n^{A-\mathrm{LDA}}(x) := \begin{cases} 1 & \text{if } \log\frac{\hat\pi_1^A}{\hat\pi_2^A} + \bigl(Ax - \frac{\hat\mu_1^A+\hat\mu_2^A}{2}\bigr)^T\hat\Omega^A(\hat\mu_1^A-\hat\mu_2^A) \geq 0; \\ 2 & \text{otherwise.}\end{cases} \tag{2.5}$$

Here, $\hat\pi_r^A := n^{-1}\sum_{i=1}^n \mathbb{1}_{\{Y_i^A = r\}}$, $\hat\mu_r^A := n^{-1}\sum_{i=1}^n AX_i\mathbb{1}_{\{Y_i^A=r\}}$,

$$\hat\Sigma^A := \frac{1}{n-2}\sum_{i=1}^n\sum_{r=1}^2 (AX_i - \hat\mu_r^A)(AX_i-\hat\mu_r^A)^T\mathbb{1}_{\{Y_i^A=r\}}$$

and $\hat\Omega^A := (\hat\Sigma^A)^{-1}$.

Write Φ for the standard normal distribution function. Under the normal model specified above, the test error of the LDA classifier can be written as

$$\mathcal{L}_n^A = \pi_1\Phi\left(\frac{\log\frac{\hat\pi_2^A}{\hat\pi_1^A} - (\hat\delta^A)^T\hat\Omega^A(\bar{\hat\mu}^A - \mu_1^A)}{\sqrt{(\hat\delta^A)^T\hat\Omega^A\Sigma^A\hat\Omega^A\hat\delta^A}}\right) + \pi_2\Phi\left(\frac{\log\frac{\hat\pi_1^A}{\hat\pi_2^A} + (\hat\delta^A)^T\hat\Omega^A(\bar{\hat\mu}^A - \mu_2^A)}{\sqrt{(\hat\delta^A)^T\hat\Omega^A\Sigma^A\hat\Omega^A\hat\delta^A}}\right),$$

where $\hat\delta^A := \hat\mu_2^A - \hat\mu_1^A$ and $\bar{\hat\mu}^A := (\hat\mu_1^A + \hat\mu_2^A)/2$. Okamoto (1963) studied the excess risk of the LDA classifier in an asymptotic regime in which d is fixed as n diverges. In fact, he considered the very slightly different discriminant analysis data generating model, in which the training sample sizes from each population are assumed to be known in advance, so that without loss of generality, we may assume that $Y_1 = \ldots = Y_{n_1} = 1$ and $Y_{n_1+1} = \ldots = Y_n = 2$, while $X_i|Y_i = r \sim N_p(\mu_r, \Sigma)$, as before. Specialising his

results for simplicity to the case where n is even and $n_1 = n_2$, Okamoto (1963) showed that using the LDA classifier (2.5) with $A = A^*$, $\hat{\pi}_1^{A^*} = n_1/n$ and $\hat{\pi}_2^{A^*} = n_2/n$ yields

$$\mathbb{E}(\mathcal{L}_n^{A^*}) - R^{A^*-\mathrm{Bayes}} = R(\hat{C}_n^{A^*}) - R^{A^*-\mathrm{Bayes}} = \frac{d}{n}\phi\Big(-\frac{\Delta}{2}\Big)\Big(\frac{\Delta}{4} + \frac{d-1}{d\Delta}\Big)\{1+O(n^{-1})\} \quad (2.6)$$

as $n \to \infty$, where $\Delta := \|\Sigma^{-1/2}(\mu_1 - \mu_2)\| = \|(\Sigma^{A^*})^{-1/2}(\mu_1^{A^*} - \mu_2^{A^*})\|$.

It remains to control the errors ϵ_n and $\epsilon_n^{A^*}$ in Theorem 2.5. For the LDA classifier, we consider the resubstitution estimator

$$\hat{L}_n^A := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{\hat{C}_n^{A-\mathrm{LDA}}(X_i) \neq Y_i\}}. \quad (2.7)$$

Devroye and Wagner (1976) provided a Vapnik–Chervonenkis bound for $\hat{L}_n^A$ under no assumptions on the underlying data generating mechanism: for every $n \in \mathbb{N}$ and $\epsilon > 0$,

$$\sup_{A \in \mathcal{A}} \mathbb{P}(|\mathcal{L}_n^A - \hat{L}_n^A| > \epsilon) \leq 8n^d e^{-n\epsilon^2/32}; \quad (2.8)$$

see also Devroye et al. (1996, Theorem 23.1). We can then conclude that

$$\begin{aligned}\mathbb{E}(|\epsilon_n^{A^*}|) \leq \mathbb{E}\big\{(\mathcal{L}_n^{A^*} - \hat{L}_n^{A^*})^2\big\}^{1/2} &\leq \inf_{\epsilon_0 \in (0,1)} \left\{\epsilon_0 + 8n^d \int_{\epsilon_0}^1 e^{-ns/32}\, ds\right\}^{1/2} \\ &\leq 8\sqrt{\frac{d\log n + 3\log 2 + 1}{2n}}. \end{aligned} \quad (2.9)$$

The more difficult term to deal with is $\mathbb{E}(|\epsilon_n|) = \mathbb{E}\big\{\big|\mathbb{E}(\mathcal{L}_n^{A_1} - \hat{L}_n^{A_1}|\mathcal{T}_n)\big|\big\} \leq \mathbb{E}\big|\mathcal{L}_n^{A_1} - \hat{L}_n^{A_1}\big|$. In this case, the bound (2.8) cannot be applied directly; since A_1 depends on the training data, the pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ are no longer independent conditional on A_1. Nevertheless, since $A_{1,1}, \ldots, A_{1,B_2}$ are independent of $\mathcal{T}_n$, we still have that

$$\begin{aligned}&\mathbb{P}\Big\{\max_{b_2=1,\ldots,B_2} |\mathcal{L}_n^{A_{1,b_2}} - \hat{L}_n^{A_{1,b_2}}| > \epsilon \;\Big|\; A_{1,1}, \ldots, A_{1,B_2}\Big\} \\ &\qquad \leq \sum_{b_2=1}^{B_2} \mathbb{P}\big\{|\mathcal{L}_n^{A_{1,b_2}} - \hat{L}_n^{A_{1,b_2}}| > \epsilon \;\big|\; A_{1,b_2}\big\} \leq 8n^d B_2 e^{-n\epsilon^2/32}. \end{aligned}$$

We can therefore conclude by almost the same argument as that leading to (2.9) that

$$\mathbb{E}(|\epsilon_n|) \leq \mathbb{E}\Big\{\max_{b_2=1,\ldots,B_2} (\mathcal{L}_n^{A_{1,b_2}} - \hat{L}_n^{A_{1,b_2}})^2\Big\}^{1/2} \leq 8\sqrt{\frac{d\log n + 3\log 2 + \log B_2 + 1}{2n}}. \quad (2.10)$$

Note that none of the bounds (2.6), (2.9) and (2.10) depend on the original data dimension p. Moreover, substituting the bound (2.10) into Theorem 2.5 reveals a trade-off in the choice of B_2 when using LDA as the base classifier. Choosing B_2 to be

large gives us a good chance of finding a projection with a small estimate of test error, but we may incur a small overfitting penalty as reflected by (2.10).

2.4.2 Quadratic Discriminant Analysis

Quadratic Discriminant Analysis (QDA) is designed to handle situations where the class-conditional covariance matrices are unequal. Recall that when $X|Y = r \sim N_p(\mu_r, \Sigma_r)$, and $\pi_r := \mathbb{P}(Y = r)$, for $r = 1, 2$, the Bayes decision boundary is given by $\{x \in \mathbb{R}^p : \Delta(x; \pi_1, \mu_1, \mu_2, \Sigma_1, \Sigma_2) = 0\}$, where

$$\Delta(x; \pi_1, \mu_1, \mu_2, \Sigma_1, \Sigma_2) := \log \frac{\pi_1}{\pi_2} - \frac{1}{2} \log\Big(\frac{\det \Sigma_1}{\det \Sigma_2}\Big) - \frac{1}{2} x^T (\Sigma_1^{-1} - \Sigma_2^{-1}) x + x^T (\Sigma_1^{-1}\mu_1 - \Sigma_2^{-1}\mu_2) - \frac{1}{2}\mu_1^T \Sigma_1^{-1}\mu_1 + \frac{1}{2}\mu_2^T \Sigma_2^{-1}\mu_2.$$

In QDA, π_r, μ_r and Σ_r are estimated by their sample versions. If $p \geq \min(n_1, n_2)$, where $n_r := \sum_{i=1}^n \mathbb{1}_{\{Y_i = r\}}$ is the number of training sample observations from the rth class, then at least one of the sample covariance matrix estimates is singular, and QDA cannot be used directly. Nevertheless, we can still choose $d < \min\{n_1, n_2\}$ and use QDA as the base classifier in a random projection ensemble. Specifically, we can set

$$\hat{C}_n^A(x) = \hat{C}_n^{A-\mathrm{QDA}}(x) := \begin{cases} 1 & \text{if } \Delta(x; \hat{\pi}_1^A, \hat{\mu}_1^A, \hat{\mu}_2^A, \hat{\Sigma}_1^A, \hat{\Sigma}_2^A) \geq 0; \\ 2 & \text{otherwise,} \end{cases}$$

where $\hat{\pi}_r^A$, $\hat{\Sigma}_r^A$ and $\hat{\mu}_r^A$ were defined in Section 2.4.1, and where

$$\hat{\Sigma}_r^A := \frac{1}{n_r - 1} \sum_{i : Y_i^A = r} (AX_i - \hat{\mu}_r^A)(AX_i - \hat{\mu}_r^A)^T$$

for $r = 1, 2$. Unfortunately, analogous theory to that presented in Section 2.4.1 does not appear to exist for the QDA classifier (unlike for LDA, the risk does not have a closed form). Nevertheless, we found in our simulations presented in Section 2.6 that the QDA random projection ensemble classifier can perform very well in practice. In this case, we estimate the test errors using the leave-one-out method given by

$$\hat{L}_n^A := \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{\hat{C}_{n,i}^A(X_i) \neq Y_i\}}, \tag{2.11}$$

where $\hat{C}_{n,i}^A$ denotes the classifier $\hat{C}_n^A$, trained without the ith pair, i.e. based on $\mathcal{T}_n^A \setminus \{X_i^A, Y_i^A\}$. For a method like QDA that involves estimating more parameters than LDA, we found that the leave-one-out estimator was less susceptible to overfitting than the resubstitution estimator.

2.4.3 The k-nearest neighbour classifier

The k-nearest neighbour classifier ($k$nn), first proposed by Fix and Hodges (1951), is a nonparametric method that classifies the test point $x \in \mathbb{R}^p$ according to a majority vote over the classes of the k nearest training data points to x. The enormous popularity of the $k$nn classifier can be attributed partly due to its simplicity and intuitive appeal; however, it also has the attractive property of being universally consistent: for every distribution P, as long as $k \to \infty$ and $k/n \to 0$, the risk of the $k$nn classifier converges to the Bayes risk (Devroye et al., 1996, Theorem 6.4).

Hall, Park and Samworth (2008) derived the rate of convergence of the excess risk of the k-nearest neighbour classifier. Under regularity conditions, the optimal choice of k, in terms of minimising the excess risk, is $O(n^{4/(p+4)})$, and the rate of convergence of the excess risk with this choice is $O(n^{-4/(p+4)})$. Thus, in common with other nonparametric methods, there is a 'curse of dimensionality' effect that means the classifier typically performs poorly in high dimensions. Samworth (2012a,b) found the optimal way of assigning decreasing weights to increasingly distant neighbours, and quantified the asymptotic improvement in risk over the unweighted version, but the rate of convergence remains the same.

We can use the $k$nn classifier as the base classifier for a random projection ensemble as follows: let $\mathcal{T}_n^A := \{(Z_1^A, Y_1^A), \ldots, (Z_n^A, Y_n^A)\}$, where $Z_i^A := AX_i$ and $Y_i^A := Y_i$. Given $z \in \mathbb{R}^d$, let $(Z_{(1)}^A, Y_{(1)}^A), \ldots, (Z_{(n)}^A, Y_{(n)}^A)$ be a re-ordering of the training data such that $\|Z_{(1)}^A - z\| \leq \ldots \leq \|Z_{(n)}^A - z\|$, with ties split at random. Now define

$$\hat{C}_n^A(x) = \hat{C}_n^{A-k\mathrm{nn}}(x) := \begin{cases} 1 & \text{if } \hat{S}_n^A(Ax) \geq 1/2; \\ 2 & \text{otherwise,} \end{cases}$$

where $\hat{S}_n^A(z) := k^{-1} \sum_{i=1}^k \mathbb{1}_{\{Y_{(i)}^A = 1\}}$. The theory described in the previous paragraph can be applied to show that, under regularity conditions, $\mathbb{E}(\mathcal{L}_n^{A^*}) - R(C^{A^*-\mathrm{Bayes}}) = O(n^{-4/(d+4)})$.

Once again, a natural estimate of the test error in this case is the leave-one-out estimator defined in (2.11), where we use the same value of k on the leave-one-out datasets as for the base classifier, and where distance ties are split in the same way as for the base classifier. For this estimator, Devroye and Wagner (1979) showed that for every $n \in \mathbb{N}$,

$$\sup_{A \in \mathcal{A}} \mathbb{E}\{(\hat{L}_n^A - \mathcal{L}_n^A)^2\} \leq \frac{1}{n} + \frac{24k^{1/2}}{n\sqrt{2\pi}};$$

see also Devroye et al. (1996, Chapter 24). It follows that

$$\mathbb{E}(|\epsilon_n^{A^*}|) \leq \Big(\frac{1}{n} + \frac{24k^{1/2}}{n\sqrt{2\pi}}\Big)^{1/2} \leq \frac{1}{n^{1/2}} + \frac{2\sqrt{3}k^{1/4}}{n^{1/2}\sqrt{\pi}}.$$

Devroye and Wagner (1979) also provided a tail bound analogous to (2.8) for the leave-one-out estimator. Arguing very similarly to Section 2.4.1, we can deduce that under no conditions on the data generating mechanism,

$$\mathbb{E}(|\epsilon_n|) \leq 3\{4(3^d+1)\}^{1/3}\left\{\frac{k(1+\log B_2+3\log 2)}{n}\right\}^{1/3}.$$

2.5 Practical considerations

2.5.1 Choice of α

We now discuss the choice of the voting threshold α in (2.2). The expression for the test error of the infinite-simulation random projection ensemble classifier given in (2.3) suggests the 'oracle' choice

$$\alpha^* \in \operatorname*{argmin}_{\alpha' \in [0,1]} \big[\pi_1 G_{n,1}(\alpha') + \pi_2\{1 - G_{n,2}(\alpha')\}\big]. \tag{2.12}$$

Note that, if assumption **(A.1)** holds and $\alpha^* \in (0,1)$ then $\pi_1 g_{n,1}(\alpha^*) = \pi_2 g_{n,2}(\alpha^*)$ and in Theorem 2.1, we have

$$\gamma_n(\alpha^*) = \frac{1}{2}\alpha^*(1-\alpha^*)\{\pi_1 \dot{g}_{n,1}(\alpha^*) - \pi_2 \dot{g}_{n,2}(\alpha^*)\}.$$

Of course, α^* cannot be used directly, because we do not know $G_{n,1}$ and $G_{n,2}$ (and we may not know π_1 and π_2 either). Nevertheless, for the LDA base classifier we can estimate $G_{n,r}$ using

$$\hat{G}_{n,r}(t) := \frac{1}{n_r}\sum_{i:Y_i=r} \mathbb{1}_{\{\hat{\nu}_n(X_i)<t\}}$$

for $r = 1, 2$. For the QDA and k-nearest neighbour base classifiers, we use the leave-one-out-based estimate $\tilde{\nu}_n(X_i) := B_1^{-1}\sum_{b_1=1}^{B_1} \mathbb{1}_{\{\hat{C}_{n,i}^{A_{b_1}}(X_i)=1\}}$ in place of $\hat{\nu}_n(X_i)$. We also estimate π_r by $\hat{\pi}_r := n^{-1}\sum_{i=1}^n \mathbb{1}_{\{Y_i=r\}}$, and then set the cut-off in (2.2) as

$$\hat{\alpha} \in \operatorname*{argmin}_{\alpha' \in [0,1]} \big[\hat{\pi}_1 \hat{G}_{n,1}(\alpha') + \hat{\pi}_2\{1 - \hat{G}_{n,2}(\alpha')\}\big]. \tag{2.13}$$

Since empirical distribution functions are piecewise constant, the objective function in (2.13) does not have a unique minimum, so we choose $\hat{\alpha}$ to be the average of the smallest and largest minimisers. An attractive feature of the method is that $\{\hat{\nu}_n(X_i) : i = 1, \ldots, n\}$ (or $\{\tilde{\nu}_n(X_i) : i = 1, \ldots, n\}$ in the case of QDA and $k$nn) are already calculated in order to choose the projections, so calculating $\hat{G}_{n,1}$ and $\hat{G}_{n,2}$ carries negligible extra computational cost.

Figures 2.2 and 2.3 illustrate $\hat{\pi}_1\hat{G}_{n,1}(\alpha') + \hat{\pi}_2\{1 - \hat{G}_{n,2}(\alpha')\}$ as an estimator of

$\pi_1 G_{n,1}(\alpha') + \pi_2\{1 - G_{n,2}(\alpha')\}$, for different base classifiers as well as different values of n and π_1. Here, a very good approximation to the estimand was obtained using an independent data set of size 5000. Unsurprisingly, the performance of the estimator improves as n increases, but the most notable feature of these plots is the fact that for all classifiers and even for small sample sizes, $\hat{\alpha}$ is an excellent estimator of α^*, and may be a substantial improvement on the naive choice $\hat{\alpha} = 1/2$ (which may result in a classifier that assigns every point to a single class).

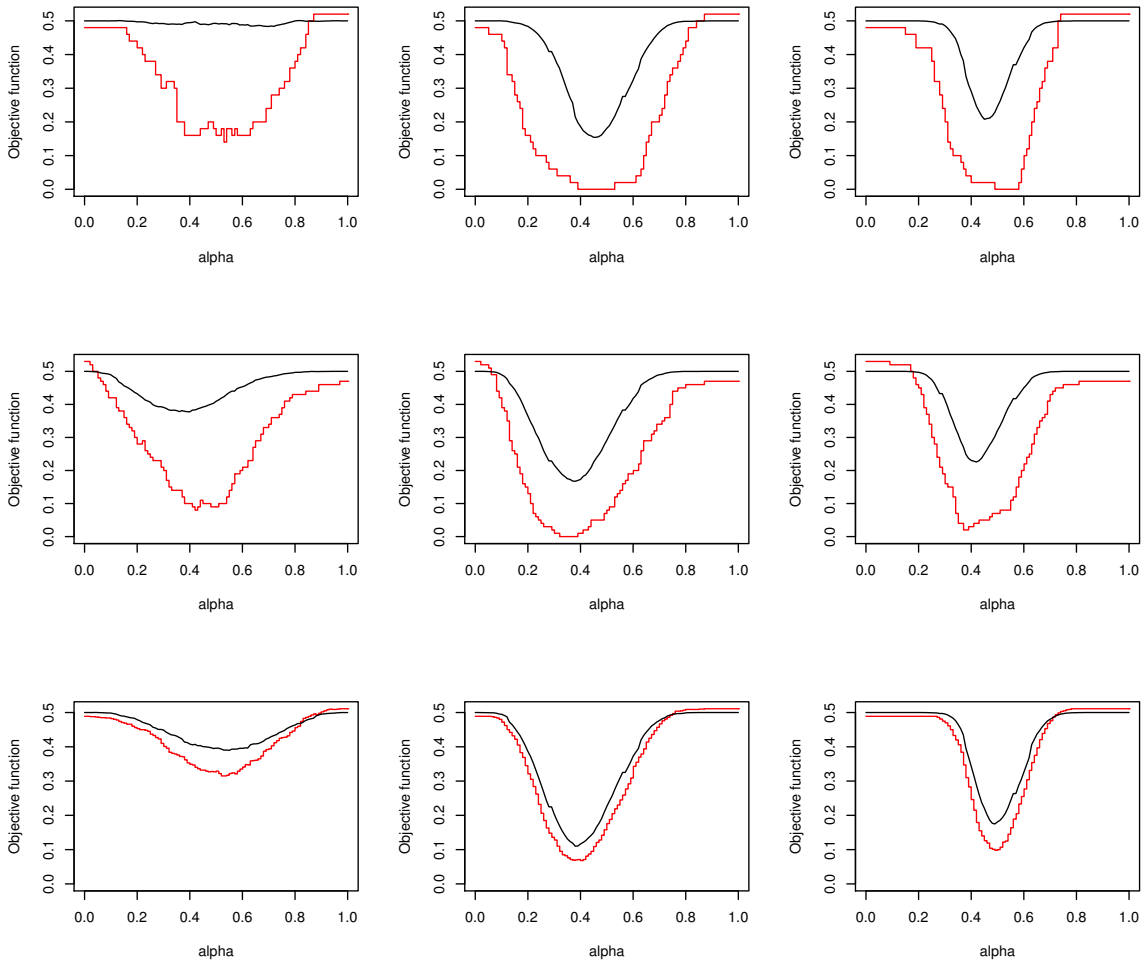


Figure 2.2: $\pi_1 G_{n,1}(\alpha') + \pi_2\{1 - G_{n,2}(\alpha')\}$ *in* (2.12) *(black) and* $\hat{\pi}_1 \hat{G}_{n,1}(\alpha') + \hat{\pi}_2\{1 - \hat{G}_{n,2}(\alpha')\}$ *(red) for the LDA (left), QDA (middle) and knn (right) base classifiers after projecting for one training data set of size* $n = 50$ *(top),* 100 *(middle) and* 1000 *(bottom) from Model 1. Here,* $\pi_1 = 0.5$, $p = 50$ *and* $d = 2$.

2.5.2 Choice of d

We want to choose d as small as possible in order to obtain the best possible performance bounds as described in Section 2.4 above. This also reduces the computational

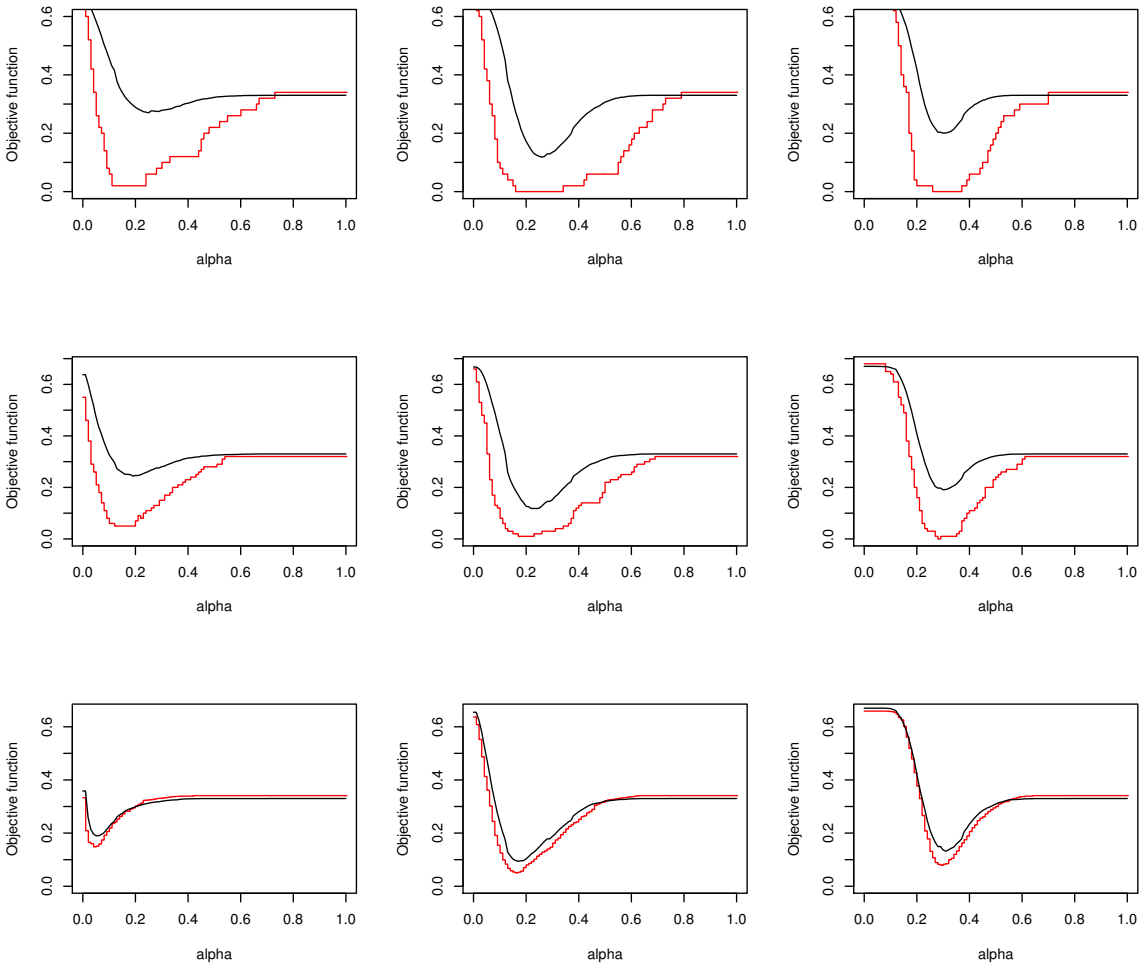


Figure 2.3: $\pi_1 G_{n,1}(\alpha') + \pi_2\{1 - G_{n,2}(\alpha')\}$ *in* (2.12) *(black) and* $\hat{\pi}_1 \hat{G}_{n,1}(\alpha') + \hat{\pi}_2\{1 - \hat{G}_{n,2}(\alpha')\}$ *(red) for the LDA (left), QDA (middle) and knn (right) base classifiers after projecting for one training data set of size* $n = 50$ *(top), 100 (middle) and 1000 (bottom) from Model 1. Here,* $\pi_1 = 0.33$*,* $p = 50$ *and* $d = 2$*.*

cost. However, the performance bounds rely on assumption **(A.3)**, whose strength decreases as d increases, so we want to choose d large enough that this condition holds (at least approximately).

In Section 2.6 we see that the random projection ensemble method is quite robust to the choice of d. Nevertheless, in some circumstances it may be desirable to have an automatic choice. As one way to do this, suppose that we wish to choose d from a set $\mathcal{D} \subseteq \{1, \ldots, p\}$. For each $d \in \mathcal{D}$, generate independent and identically distributed projections $\{A_{d,b_1,b_2} : b_1 = 1, \ldots, B_1, b_2 = 1, \ldots, B_2\}$ according to Haar measure on $\mathcal{A}_{d\times p}$. For each $d \in \mathcal{D}$ and for $b = 1, \ldots, B_1$, we can then set

$$A_{d,b_1} := A_{d,b_1,b_2^*(b_1)},$$

where $b_2^*(b_1) := \text{sargmin}_{b_2 \in \{1,\ldots,B_2\}} \hat{L}_n^{A_{d,b_1,b_2}}$. Finally, we can select

$$\hat{d} := \underset{d \in \mathcal{D}}{\text{sargmin}} \frac{1}{B_1} \sum_{b_1=1}^{B_1} \hat{L}_n^{A_{d,b_1}}.$$

In Figures 2.4 and 2.5 below we present the empirical distribution functions of $\{\hat{L}_n^{A_{d,b_1}}\}_{b_1=1}^{B_1}$, where $d \in \{2,3,4,5\}$, for one training dataset from Model 1 (described in Section 2.6.1), and Model 3 (described in Section 2.6.1). In each case, we set $\pi_1 = 1/2$, $n = 100$, $p = 50$ and $B_1 = B_2 = 100$.

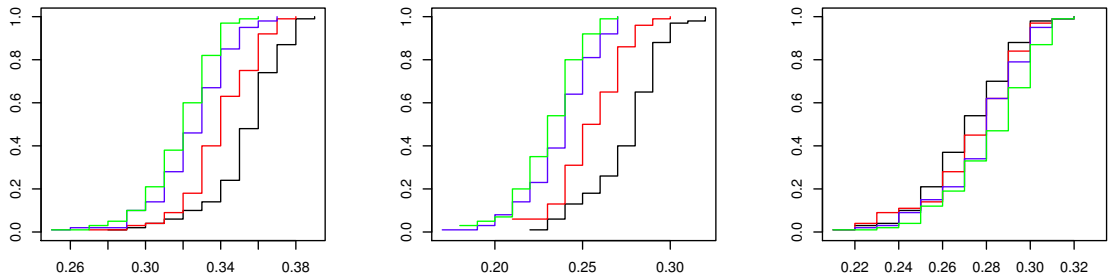


Figure 2.4: *Empirical distribution functions of the test error estimates* $\{\hat{L}_n^{A_{d,b_1}}\}_{b_1=1}^{B_1}$ *for the LDA (left), QDA (middle) and knn (right) base classifiers after projecting for Model 1,* $\pi_1 = 1/2$, $n = 100$, $p = 50$, *and* $d = 2$ *(black), 3 (red), 4 (blue) and 5 (green).*

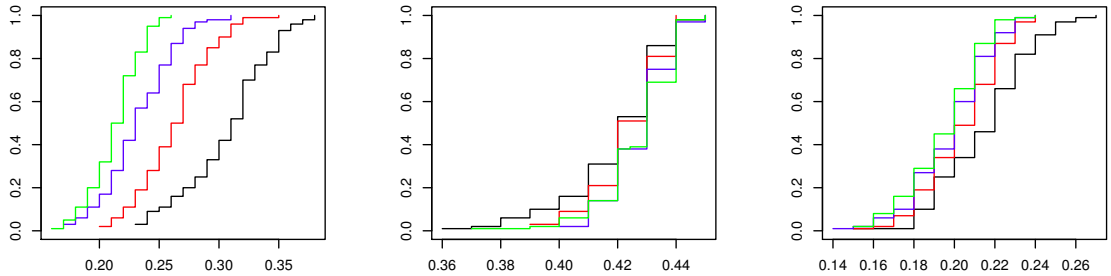


Figure 2.5: *Empirical distribution functions of the test error estimates* $\{\hat{L}_n^{A_{d,b_1}}\}_{b_1=1}^{B_1}$ *for the LDA (left), QDA (middle) and knn (right) base classifiers after projecting for Model 2,* $\pi_1 = 1/2$, $n = 100$, $p = 50$, *and* $d = 2$ *(black), 3 (red), 4 (blue) and 5 (green).*

Figures 2.4 and 2.5 do not suggest great differences in performance for different choices of d, especially for the QDA and $k$nn base classifiers. For the LDA classifier, it appears, particularly for Model 2, that projecting into a slightly larger dimensional space is preferable, and indeed this appears to be the case from the relevant entry of Table 2.2 below.

The ideas presented here may also be used to decide between two different base classifiers. For example, comparing the green lines across different panels of Figure 2.4, we see that for Model 1 and $d = 5$, we might expect the best results with the QDA base classifier, and indeed this is confirmed by the simulation results in Table 2.1 below.

2.5.3 Choice of B_1 and B_2

In order to minimise the third term in the bound in Theorem 2.5, we should choose B_1 to be as large as possible. The constraint, of course, is that the computational cost of the random projection classifier scales linearly with B_1. The choice of B_2 is more subtle; while the fourth term in the bound in Theorem 2.5 decreases as B_2 increases, we saw in Section 2.4 that upper bounds on $\mathbb{E}(|\epsilon_n|)$ may increase with B_2. In principle, we could try to use the expressions given in Theorem 2.5 and Section 2.4 to choose B_2 to minimise the overall upper bound on $\mathcal{L}(\hat{C}_n^{\mathrm{RP}}) - R(C^{\mathrm{Bayes}})$. Although β_0, β and ρ are unknown to the practitioner, these could be estimated based on the empirical distribution of $\{\hat{L}_n^{A_b}\}_{b=1}^B$, where $\{A_b\}_{b=1}^B$ are independent projections drawn according to Haar measure. In practice, however, we found that an involved approach such as this was unnecessary, and that the ensemble method was robust to the choice of B_1 and B_2. In all of our simulations, we set $B_1 = B_2 = 100$.

2.6 Empirical analysis

In this section, we assess the empirical performance of the random projection ensemble classifier. Throughout, RP-LDA$_d$, RP-QDA$_d$, and RP-$k$nn$_d$ denote the random projection ensemble classifier with LDA, QDA, and $k$nn base classifiers, respectively; the subscript d refers to the dimension of the image space of the projections. For comparison, we present the results of the related Random Forests (RF) classifier (Breiman, 2001); and the widely used methods, Support Vector Machines (SVM) (Cortes and Vapnik, 1995) and Gaussian Process (GP) classifiers (Williams and Barber, 1998). We also present results for several state-of-the-art methods for high-dimensional classification, namely Penalized LDA (PenLDA) (Witten and Tibshirani, 2011), Nearest Shrunken Centroids (NSC) (Tibshirani et al., 2003), Shrunken Centroids Regularized Discriminant Analysis (SCRDA) (Guo, Hastie and Tibshirani, 2007), and Independence Rules (IR) (Bickel and Levina, 2004), as well as for the base classifier applied in the original space.

For the standard $k$nn classifier, we chose k via leave-one-out cross validation. The Random Forest used was an ensemble of 1000 trees, with $\lceil\sqrt{p}\rceil$ components randomly selected when training each tree. This method was implemented using the `randomForest` package (Liaw and Wiener, 2014). For the Radial SVM, we used the

reproducing basis kernel $K(u,v) = \exp(-\frac{1}{p}\|u-v\|^2)$. Both SVM classifiers were implemented using the `svm` function in the `e1071` package (Meyer et. al., 2015). For the GP classifier, we used a radial basis function, with the hyperparameter chosen via the automatic method in the `guasspr` function in the `kernlab` package (Karatzoglou et al., 2015). The tuning parameters for the other methods were chosen using the default settings in the corresponding `R` packages `PenLDA` (Witten, 2011), `NSC` (Hastie et al., 2015) and `SCRDA` (Islam and Mcleod, 2015), namely 6-fold, 10-fold and 10-fold cross validation, respectively.

2.6.1 Simulated examples

In each of the simulated examples below, we used $n \in \{50, 100, 200\}$, $d = 2, 5$ and $B_1 = B_2 = 100$, the experiment was repeated 100 times and the risk was estimated on an independent test set of size 1000. For all except Model 5, we take $p = 50$, investigate two different values of the prior probability, and use Haar projections. In Model 5, where the class label is deterministic, we investigate two different values of p, namely, 20 and 50. Moreover, in this example, we use axis-aligned projections, see Section 2.7 for further discussion. The mean and standard error of the risk estimates are presented in Tables 2.1 - 2.5. In each case, we highlight the method achieving the best performance.

Independent features

Model 1: Here, P_1 is the distribution of p independent components, each with a standard Laplace distribution, while $P_2 = N_p(\mu, I_{p\times p})$, with $\mu = \frac{1}{8}(1, \ldots, 1)^T$. We set $p = 50$.

In Model 1, the class boundaries are non-linear and, in fact, assumption **(A.3)** is not satisfied for any $d < p$. Nevertheless, in Table 2.1 we see that the random projection versions outperform their respective vanilla counterparts (when these are tractable) as well as nearly all of the other methods implemented. The only exception is that the radial SVM performs well when the priors are equal and the sample size is large, surprisingly the SVM classifier doesn't cope well with the change in prior. The RP-QDA classifiers perform especially well; in particular, they are able to cope better with the non-linearity of the class boundaries than the RP-LDA classifiers. There is little difference between the performance of the $d = 2$ and $d = 5$ versions of the random projection classifiers.

Table 2.1: *Misclassification rates multiplied by 100 (with standard errors as subscripts) for Model 1, with smallest in bold. *N/A: not available due to singular covariance estimates.*

	$\pi_1 = 0.5$, Bayes risk = 4.91			$\pi_1 = 0.33$, Bayes risk = 4.09		
n	50	100	200	50	100	200
RP-LDA$_2$	$43.99_{0.64}$	$41.99_{0.58}$	$41.14_{0.55}$	$29.83_{0.51}$	$26.70_{0.33}$	$23.76_{0.25}$
RP-LDA$_5$	$43.95_{0.44}$	$42.10_{0.39}$	$41.15_{0.37}$	$33.11_{0.52}$	$30.03_{0.36}$	$26.55_{0.24}$
LDA	N/A	$44.17_{0.27}$	$41.88_{0.20}$	N/A	$37.64_{0.31}$	$33.75_{0.20}$
RP-QDA$_2$	$19.74_{0.43}$	$15.65_{0.24}$	$13.60_{0.15}$	$\mathbf{17.64}_{0.50}$	$13.37_{0.27}$	$11.88_{0.21}$
RP-QDA$_5$	$\mathbf{19.05}_{0.42}$	$\mathbf{14.05}_{0.24}$	$11.75_{0.14}$	$18.06_{0.46}$	$\mathbf{12.86}_{0.31}$	$\mathbf{10.64}_{0.16}$
QDA	N/A	N/A	$39.93_{0.29}$	N/A	N/A	$33.05_{0.16}$
RP-$k$nn$_2$	$26.19_{0.37}$	$20.96_{0.23}$	$18.56_{0.14}$	$23.58_{0.34}$	$20.02_{0.22}$	$16.58_{0.21}$
RP-$k$nn$_5$	$27.41_{0.35}$	$21.30_{0.24}$	$17.48_{0.15}$	$24.64_{0.31}$	$19.33_{0.28}$	$16.15_{0.20}$
$k$nn	$48.81_{0.18}$	$48.68_{0.19}$	$48.29_{0.18}$	$32.69_{0.17}$	$32.50_{0.16}$	$32.64_{0.15}$
RF	$42.75_{0.31}$	$35.63_{0.25}$	$26.25_{0.23}$	$31.65_{0.20}$	$28.21_{0.20}$	$22.92_{0.21}$
Radial SVM	$38.38_{1.36}$	$15.81_{0.54}$	$\mathbf{10.81}_{0.13}$	$32.03_{0.46}$	$30.48_{0.48}$	$22.27_{0.72}$
Linear SVM	$45.24_{0.24}$	$44.13_{0.26}$	$42.44_{0.22}$	$36.50_{0.39}$	$35.84_{0.29}$	$31.82_{0.18}$
Radial GP	$47.14_{0.35}$	$44.32_{0.43}$	$39.86_{0.34}$	$32.79_{0.17}$	$32.67_{0.16}$	$32.66_{0.16}$
PenLDA	$44.40_{0.27}$	$42.60_{0.25}$	$41.05_{0.20}$	$33.19_{0.33}$	$32.61_{0.25}$	$31.31_{0.17}$
NSC	$46.51_{0.33}$	$44.60_{0.39}$	$43.03_{0.39}$	$31.76_{0.21}$	$31.13_{0.17}$	$31.65_{0.18}$
SCRDA	$46.76_{0.31}$	$44.55_{0.38}$	$42.55_{0.37}$	$33.56_{0.37}$	$32.52_{0.23}$	$31.94_{0.18}$
IR	$43.87_{0.24}$	$42.25_{0.25}$	$40.55_{0.18}$	$35.04_{0.34}$	$36.26_{0.28}$	$36.48_{0.23}$

t-distributed features

Model 2: Here, $X|\{Y = r\} = \mu_r + \frac{Z}{\sqrt{U/\nu_r}}$, where $Z \sim N_p(0, \Sigma_r)$ independent of $U \sim \chi^2_{\nu_r}$. Thus, P_r is the multivariate t-distribution centred at μ_r, with ν_r degrees of freedom and shape parameter Σ_r. We set $p = 50$, $\mu_1 = 0$, $\mu_2 = 2(1, \ldots, 1, 0, \ldots, 0)^T$, where μ_2 has 5 non-zero components, $\nu_1 = 1$, $\nu_2 = 2$, $\Sigma_1 = I_{p\times p}$ and $\Sigma_2 = (\Sigma_{j,k})$, where $\Sigma_{j,j} = 1$, $\Sigma_{j,k} = 0.5$ if $\max(j, k) \leq 5$ and $j \neq k$, $\Sigma_{j,k} = 0$ otherwise.

Model 2 explores the effect of heavy tails and the presence of correlation between the features. Again, assumption **(A.3)** is not satisfied for any $d < p$. We see in Table 2.2 that both the RP-$k$nn and Random Forest classifiers perform well. There is little difference between the $d = 2$ and $d = 5$ versions of the random projection ensembles. The RP-QDA classifiers perform poorly here because the heavy-tailed distributions leads to poor mean and covariance matrix estimates.

Multi-modal features

Model 3: Here, $X|\{Y = 1\} \sim \frac{1}{2}N_p(\mu_1, \Sigma_1) + \frac{1}{2}N_p(-\mu_1, \Sigma_1)$ and $X|\{Y = 2\}$ has p independent components, the first five of which are standard Cauchy and the remaining $p - 5$ of which are standard normal. We set $p = 50$, $\mu_1 = (1, \ldots, 1, 0, \ldots, 0)^T$, where μ_1 has 5 non-zero components, and $\Sigma_1 = I_{p\times p}$.

Table 2.2: *Misclassification rates for Model 2.*

	$\pi_1 = 0.5$, Bayes risk = 10.07			$\pi_1 = 0.75$, Bayes risk = 6.67		
n	50	100	200	50	100	200
RP-LDA$_2$	$23.86_{0.91}$	$21.74_{0.81}$	$22.90_{1.06}$	$29.43_{0.48}$	$26.69_{0.34}$	$25.48_{0.19}$
RP-LDA$_5$	$21.14_{0.52}$	$18.23_{0.35}$	$17.50_{0.43}$	$31.29_{0.45}$	$30.26_{0.43}$	$27.26_{0.31}$
LDA	N/A	$26.49_{0.28}$	$21.74_{0.22}$	N/A	$23.35_{0.26}$	$20.15_{0.23}$
RP-QDA$_2$	$32.97_{0.87}$	$33.41_{1.03}$	$38.89_{0.71}$	$18.97_{0.79}$	$22.95_{1.08}$	$27.34_{1.07}$
RP-QDA$_5$	$34.91_{0.69}$	$37.35_{0.75}$	$40.90_{0.40}$	$20.25_{0.74}$	$28.78_{1.09}$	$40.56_{0.96}$
QDA	N/A	N/A	$39.95_{0.23}$	N/A	N/A	N/A
RP-$k$nn$_2$	$17.16_{0.31}$	$15.02_{0.20}$	$13.31_{0.14}$	$13.24_{0.37}$	$11.08_{0.23}$	$9.51_{0.13}$
RP-$k$nn$_5$	$16.86_{0.27}$	$15.37_{0.22}$	$13.56_{0.13}$	$\mathbf{12.17}_{0.31}$	$\mathbf{9.99}_{0.17}$	$\mathbf{9.02}_{0.12}$
$k$nn	$20.26_{0.37}$	$18.33_{0.18}$	$16.79_{0.14}$	$15.64_{0.33}$	$13.71_{0.25}$	$12.18_{0.13}$
RF	$\mathbf{16.41}_{0.34}$	$\mathbf{14.13}_{0.19}$	$\mathbf{12.42}_{0.12}$	$14.48_{0.45}$	$10.56_{0.18}$	$9.23_{0.12}$
Radial SVM	$44.36_{0.85}$	$40.08_{1.12}$	$35.88_{1.33}$	$25.29_{0.14}$	$24.75_{0.13}$	$24.99_{0.13}$
Linear SVM	$25.69_{0.89}$	$23.74_{0.80}$	$21.73_{0.81}$	$23.50_{0.41}$	$22.99_{0.36}$	$24.17_{0.33}$
Radial GP	$23.55_{0.95}$	$15.49_{0.24}$	$13.99_{0.12}$	$20.90_{0.44}$	$17.40_{0.37}$	$13.37_{0.18}$
PenLDA	$38.69_{1.08}$	$35.07_{1.24}$	$35.19_{1.23}$	$28.96_{1.19}$	$25.69_{0.33}$	$26.11_{0.19}$
NSC	$40.03_{1.19}$	$38.48_{1.24}$	$39.29_{1.19}$	$26.18_{0.21}$	$25.25_{0.14}$	$25.32_{0.13}$
SCRDA	$21.95_{0.59}$	$18.35_{0.28}$	$16.98_{0.20}$	$20.34_{0.44}$	$18.35_{0.32}$	$16.98_{0.22}$
IR	$39.24_{1.04}$	$37.02_{1.15}$	$38.06_{1.08}$	$48.45_{2.08}$	$49.10_{2.02}$	$50.93_{1.98}$

Table 2.3: *Misclassification rates for Model 3.*

	$\pi_1 = 0.5$, Bayes risk = 11.58			$\pi_1 = 0.75$, Bayes risk = 13.13		
n	50	100	200	50	100	200
RP-LDA$_2$	$43.95_{0.63}$	$42.57_{0.61}$	$41.60_{0.70}$	$23.41_{0.57}$	$22.24_{0.34}$	$22.31_{0.22}$
RP-LDA$_5$	$45.29_{0.49}$	$44.57_{0.43}$	$43.78_{0.49}$	$27.15_{0.68}$	$23.67_{0.43}$	$22.48_{0.24}$
LDA	N/A	$49.39_{0.19}$	$49.37_{0.16}$	N/A	$35.33_{0.43}$	$30.57_{0.29}$
RP-QDA$_2$	$29.78_{0.57}$	$26.34_{0.42}$	$23.47_{0.29}$	$19.36_{0.70}$	$17.19_{0.44}$	$15.36_{0.30}$
RP-QDA$_5$	$29.93_{0.64}$	$\mathbf{24.83}_{0.39}$	$\mathbf{22.20}_{0.26}$	$20.04_{0.86}$	$\mathbf{16.50}_{0.45}$	$\mathbf{14.23}_{0.28}$
QDA	N/A	N/A	$27.58_{0.34}$	N/A	N/A	N/A
RP-$k$nn$_2$	$29.43_{0.39}$	$26.48_{0.61}$	$23.57_{0.23}$	$\mathbf{19.31}_{0.37}$	$16.64_{0.26}$	$14.61_{0.20}$
RP-$k$nn$_5$	$\mathbf{29.36}_{0.38}$	$26.29_{0.29}$	$23.38_{0.19}$	$19.93_{0.40}$	$18.82_{0.36}$	$14.74_{0.18}$
$k$nn	$34.46_{0.32}$	$31.26_{0.23}$	$28.88_{0.21}$	$22.46_{0.20}$	$19.85_{0.19}$	$17.70_{0.16}$
RF	$40.49_{0.42}$	$33.51_{0.32}$	$25.02_{0.22}$	$24.50_{0.16}$	$22.06_{0.18}$	$17.52_{0.19}$
Radial SVM	$48.87_{0.36}$	$49.66_{0.19}$	$48.18_{0.31}$	$25.09_{0.13}$	$24.96_{0.13}$	$24.96_{0.14}$
Linear SVM	$48.66_{0.20}$	$49.31_{0.19}$	$48.87_{0.19}$	$34.17_{0.53}$	$32.70_{0.41}$	$24.71_{0.23}$
Radial GP	$38.94_{0.48}$	$33.62_{0.34}$	$29.66_{0.25}$	$23.06_{0.14}$	$21.75_{0.14}$	$20.23_{0.15}$
PenLDA	$48.34_{0.22}$	$48.97_{0.37}$	$49.21_{0.17}$	$27.54_{0.50}$	$27.05_{0.42}$	$26.26_{0.29}$
NSC	$47.67_{0.42}$	$47.69_{0.37}$	$47.83_{0.32}$	$23.03_{0.16}$	$23.12_{0.15}$	$23.63_{0.14}$
SCRDA	$45.44_{0.52}$	$44.86_{0.55}$	$43.27_{0.51}$	$23.65_{0.42}$	$21.89_{0.27}$	$21.63_{0.19}$
IR	$48.40_{0.21}$	$48.91_{0.17}$	$49.29_{0.16}$	$32.26_{0.59}$	$35.35_{0.53}$	$38.54_{0.37}$

Model 3 is chosen to investigate a setting in which one class is multi-modal. Note that assumption **(A.3)** holds with $d = 5$; indeed, for example, the five rows of A^* may be taken to be the first five standard Euclidean basis vectors. In Table 2.3, we

see that the the random projection ensembles for each of the three base classifiers outperform their standard counterparts. The RP-$k$nn and RP-QDA classifiers are particularly effective here. Even though the Bayes decision boundary here is sparse – it only depends on the first 5 features – the PenLDA, NSC, SCRDA and IR classifiers perform poorly because the Bayes decision boundary is non-linear.

Rotated Sparse Normal

Model 4: Here, $X|\{Y=1\} \sim N_p(R\mu_1, R\Sigma_1 R^T)$ and $X|\{Y=2\} \sim N_p(R\mu_2, R\Sigma_2 R^T)$ where R is a $p \times p$ rotation matrix that was sampled once according to Haar measure, and remained fixed thereafter, and we set $p = 50$, $\mu_1 = (0, \ldots, 0)^T$, $\mu_2 = (1, 1, 1, 0, \ldots, 0)^T$. Moreover, Σ_1 and Σ_2 are block diagonal, with blocks $\Sigma_{r,1}$, and $\Sigma_{r,2}$, for $r = 1, 2$, where $\Sigma_{1,1}$ is a 3×3 matrix with diagonal entries equal to 1 and off-diagonal entries equal to $1/2$, and $\Sigma_{2,1} = \Sigma_{1,1} + I_{3\times 3}$. In both cases $\Sigma_{r,2}$ is a $(p-3) \times (p-3)$ matrix, with diagonal entries equal to 1 and off-diagonal entries equal to $1/2$.

Table 2.4: *Misclassification rates for Model 4.*

	$\pi_1 = 0.5$, Bayes risk = 11.83			$\pi_1 = 0.75$, Bayes risk = 7.21		
n	50	100	200	50	100	200
RP-LDA$_2$	$35.34_{0.34}$	$32.50_{0.22}$	$30.52_{0.17}$	$26.79_{0.47}$	$23.49_{0.30}$	$\mathbf{21.87}_{0.21}$
RP-LDA$_5$	$35.35_{0.28}$	$32.42_{0.18}$	$30.52_{0.14}$	$28.24_{0.54}$	$23.35_{0.28}$	$22.10_{0.21}$
LDA	N/A	$41.03_{0.26}$	$36.34_{0.21}$	N/A	$32.90_{0.36}$	$27.30_{0.23}$
RP-QDA$_2$	$\mathbf{35.31}_{0.32}$	$32.56_{0.25}$	$30.36_{0.16}$	$26.82_{0.53}$	$23.47_{0.31}$	$22.09_{0.23}$
RP-QDA$_5$	$35.68_{0.29}$	$\mathbf{32.20}_{0.23}$	$\mathbf{30.06}_{0.14}$	$28.86_{0.54}$	$23.94_{0.31}$	$22.10_{0.22}$
QDA	N/A	N/A	$44.13_{0.19}$	N/A	N/A	N/A
RP-$k$nn$_2$	$36.78_{0.35}$	$33.30_{0.21}$	$31.09_{0.17}$	$26.86_{0.47}$	$24.45_{0.29}$	$22.35_{0.21}$
RP-$k$nn$_5$	$36.58_{0.34}$	$33.10_{0.23}$	$30.77_{0.17}$	$26.18_{0.36}$	$23.78_{0.26}$	$22.19_{0.19}$
$k$nn	$40.39_{0.29}$	$38.96_{0.26}$	$37.43_{0.19}$	$26.27_{0.28}$	$25.77_{0.26}$	$25.00_{0.19}$
RF	$37.29_{0.36}$	$33.35_{0.21}$	$31.09_{0.15}$	$\mathbf{24.32}_{0.18}$	$\mathbf{23.13}_{0.18}$	$22.09_{0.17}$
Radial SVM	$46.53_{0.59}$	$40.50_{0.65}$	$32.96_{0.40}$	$25.06_{0.13}$	$25.19_{0.14}$	$24.98_{0.16}$
Linear SVM	$40.66_{0.34}$	$39.43_{0.25}$	$36.37_{0.20}$	$30.53_{0.46}$	$28.32_{0.30}$	$25.22_{0.18}$
Radial GP	$39.13_{0.47}$	$33.61_{0.25}$	$31.02_{0.16}$	$24.75_{0.14}$	$24.46_{0.15}$	$23.11_{0.17}$
PenLDA	$36.22_{0.41}$	$33.49_{0.31}$	$31.24_{0.22}$	$29.65_{0.56}$	$27.73_{0.50}$	$26.59_{0.40}$
NSC	$39.06_{0.51}$	$35.41_{0.36}$	$32.77_{0.27}$	$25.63_{0.43}$	$24.19_{0.29}$	$22.90_{0.26}$
SCRDA	$39.28_{0.53}$	$34.26_{0.40}$	$30.78_{0.19}$	$25.62_{0.32}$	$23.84_{0.24}$	$22.93_{0.20}$
IR	$36.40_{0.43}$	$33.67_{0.34}$	$31.23_{0.26}$	$32.72_{0.59}$	$31.42_{0.49}$	$30.88_{0.40}$

The setting in Model 4 is chosen so that assumption **(A.3)** holds with $d = 3$; in fact, A^* can be taken to be the first three rows of R^T. Note that, if $R = I_{p\times p}$, then the model would be sparse and we would expect the PenLDA, NSC, and SCRDA methods to perform very well. However, for a generic R, the model is not sparse and the random projection ensemble methods, which are invariant to the choice of coordinate system, typically perform as well or better (especially RP-LDA and RP-QDA). The random

forest method also performs well here. Classification is difficult in this setting, and the risks of all of the classifiers are considerably greater than the Bayes risk.

Sphere vs. Cube

Model 5: Here, the first 3 components of X, independently of the remaining $p-3$, have a uniform distribution on the unit cube $\{z \in \mathbb{R}^3 : -1 \leq z_1, z_2, z_3 \leq 1\}$. The other $p-3$ components have a $N_{p-3}(0, I_{p-3})$ distribution. Then, we set $Y = 1$ if the first three components of X lie outside the unit sphere $\{z \in \mathbb{R}^3 : z_1^2 + z_2^2 + z_3^3 \leq 1\}$, and $Y = 2$ otherwise. The prior for class 1 is $1 - \pi/6 \sim 0.48$. We conduct the experiment for $p = 20, 50$. Due to the number of noise variables, the random projection ensembles with Haar projections do not perform well here, however, since the model is sparse, axis-aligned projections – see the discussion in Section 2.7 – are able to pick up the class structure. Note that **(A.3)** holds with $d = 3$, for example A^* can be taken to be the projection that picks out the first three Euclidean basis vectors.

Table 2.5: *Misclassification rates for Model 5. *The random projection ensembles here are using axis-aligned projections.*

	$p = 20$, $\pi_1 = 0.48$			$p = 50$, $\pi_1 = 0.48$		
n	50	100	200	50	100	200
*RP-LDA$_2$	$47.16_{0.46}$	$46.64_{0.45}$	$43.74_{0.49}$	$48.34_{0.39}$	$46.14_{0.46}$	$45.05_{0.50}$
*RP-LDA$_5$	$48.54_{0.34}$	$48.92_{0.27}$	$46.96_{0.39}$	$49.20_{0.26}$	$48.94_{0.26}$	$47.90_{0.36}$
LDA	$49.63_{0.18}$	$49.60_{0.17}$	$49.30_{0.17}$	N/A	$49.62_{0.16}$	$49.80_{0.16}$
*RP-QDA$_2$	$\mathbf{33.26}_{0.46}$	$25.86_{0.45}$	$20.25_{0.37}$	$\mathbf{33.42}_{0.57}$	$27.08_{0.37}$	$24.93_{0.44}$
*RP-QDA$_5$	$35.55_{0.46}$	$25.23_{0.40}$	$\mathbf{17.09}_{0.32}$	$37.26_{0.51}$	$\mathbf{25.42}_{0.39}$	$\mathbf{18.60}_{0.32}$
QDA	$48.46_{0.23}$	$44.35_{0.15}$	$40.16_{0.18}$	N/A	N/A	$46.93_{0.17}$
*RP-$k$nn$_2$	$33.33_{0.58}$	$\mathbf{24.99}_{0.29}$	$20.60_{0.29}$	$33.78_{0.47}$	$25.61_{0.34}$	$23.19_{0.38}$
*RP-$k$nn$_5$	$43.40_{0.36}$	$33.89_{0.39}$	$21.71_{0.22}$	$47.06_{0.23}$	$39.56_{0.41}$	$25.37_{0.34}$
$k$nn	$48.97_{0.16}$	$48.70_{0.17}$	$48.35_{0.18}$	$49.73_{0.17}$	$49.40_{0.17}$	$49.15_{0.15}$
RF	$33.78_{0.39}$	$25.27_{0.30}$	$18.54_{0.24}$	$40.09_{0.37}$	$31.47_{0.39}$	$22.59_{0.31}$
Radial SVM	$48.75_{0.28}$	$47.99_{0.26}$	$46.00_{0.25}$	$49.32_{0.26}$	$48.96_{0.25}$	$48.70_{0.24}$
Linear SVM	$49.16_{0.22}$	$49.44_{0.19}$	$48.37_{0.21}$	$49.68_{0.19}$	$49.54_{0.18}$	$49.73_{0.17}$
Radial GP	$49.22_{0.22}$	$49.26_{0.19}$	$48.62_{0.17}$	$49.42_{0.23}$	$49.14_{0.21}$	$49.35_{0.19}$
PenLDA	$49.37_{0.21}$	$49.64_{0.18}$	$49.23_{0.17}$	$49.58_{0.18}$	$49.59_{0.18}$	$49.75_{0.15}$
NSC	$49.47_{0.26}$	$49.14_{0.25}$	$48.81_{0.25}$	$49.81_{0.22}$	$49.06_{0.22}$	$49.14_{0.23}$
SCRDA	$49.24_{0.34}$	$49.14_{0.25}$	$48.21_{0.30}$	$49.64_{0.22}$	$48.96_{0.22}$	$48.65_{0.32}$
IR	$49.53_{0.17}$	$49.85_{0.16}$	$49.76_{0.17}$	$49.73_{0.17}$	$49.59_{0.17}$	$49.92_{0.15}$

The setting in Model 5 is such that the Bayes risk is zero, but it represents a situation where existing classifiers perform poorly; indeed, only the random forest classifier is much better than a random guess. The random projection ensembles with axis-aligned projections and the QDA or $k$nn base classifiers perform comparatively very well. The LDA base classifier does not perform well here since the boundary is

non-linear.

2.6.2 Real data examples

In this section, we compare the classifiers discussed at the beginning of this section on four real datasets available from the UC Irvine (UCI) Machine Learning Repository (`http://archive.ics.uci.edu/ml/index.html`). In each example, we first subsample the data to obtain a training set of size $n \in \{50, 100, 200\}$, then use the remaining data (or take a subsample of size 1000 from remaining data, whichever is smaller) to form the test set. As with the simulated examples, each experiment is repeated 100 times and we present the mean and standard error of the risk estimates in Tables 2.6 and 2.7. In each case we took $B_1 = B_2 = 100$ and used Haar distributed projections.

Eye state detection

The Electroencephalogram (EEG) Eye State dataset[2] contains $p = 14$ EEG measurements for 14980 observations. The task is to use the EEG reading to determine the state of the eye. There are 8256 observations for which the eye is open (class 1), and 6723 for which the eye is closed (class 2). In this example, see Table 2.6, the random projection ensemble with $k$nn base classifier and $d = 5$ is very good, outperforming all other comparators.

Ionosphere dataset

The Ionosphere dataset[3] consists of $p = 32$ high-frequency antenna measurements for 351 observations. Observations are classified as good (class 1, size 225) or bad (class 2, size 126), depending on whether there is evidence for free electrons in the ionosphere or not. In the right panel of Table 2.6, we see that the random projection ensembles with the QDA and $k$nn base classifiers are performing very well.

Musk identification

The Musk dataset[4] consists of 1016 musk (class 1) and 5581 non-musk (class 2) molecules. The task is to classify a new molecule, based on $p = 166$ shape measurements. The results are presented in the left panel of Table 2.7. We see that all of the random projection ensemble classifiers performed very well, especially with the $k$nn base classifier.

[2] `http://archive.ics.uci.edu/ml/datasets/EEG+Eye+State`
[3] `http://archive.ics.uci.edu/ml/datasets/Ionosphere`
[4] `https://archive.ics.uci.edu/ml/datasets/Musk+(Version+2)`

Table 2.6: *Misclassification rates for the Eye State and Ionosphere datasets.*

	Eye State $p = 14$			Ionosphere $p = 32$		
n	50	100	200	50	100	200
RP-LDA$_2$	$41.64_{0.30}$	$39.84_{0.26}$	$38.70_{0.24}$	$13.12_{0.39}$	$10.75_{0.29}$	$9.68_{0.25}$
RP-LDA$_5$	$42.14_{0.36}$	$40.05_{0.28}$	$38.42_{0.24}$	$13.10_{0.31}$	$11.08_{0.21}$	$10.07_{0.24}$
LDA	$42.38_{0.37}$	$40.82_{0.29}$	$39.15_{0.26}$	$23.72_{0.38}$	$18.27_{0.25}$	$15.58_{0.29}$
RP-QDA$_2$	$40.84_{0.36}$	$38.34_{0.34}$	$36.04_{0.41}$	$9.50_{0.34}$	$6.96_{0.19}$	$5.89_{0.18}$
RP-QDA$_5$	$38.94_{0.33}$	$35.66_{0.37}$	$32.27_{0.40}$	$\mathbf{8.20}_{0.35}$	$\mathbf{6.20}_{0.18}$	$\mathbf{5.05}_{0.18}$
QDA	$39.91_{0.35}$	$33.52_{0.30}$	$29.24_{0.38}$	N/A	N/A	$14.07_{0.32}$
RP-$k$nn$_2$	$39.20_{0.35}$	$34.93_{0.29}$	$30.70_{0.24}$	$10.48_{0.32}$	$6.64_{0.20}$	$5.34_{0.17}$
RP-$k$nn$_5$	$\mathbf{37.78}_{0.36}$	$\mathbf{31.58}_{0.29}$	$\mathbf{24.47}_{0.23}$	$11.44_{0.38}$	$7.21_{0.22}$	$5.09_{0.18}$
$k$nn	$41.65_{0.36}$	$35.65_{0.36}$	$29.18_{0.23}$	$21.81_{0.72}$	$18.05_{0.45}$	$16.39_{0.33}$
RF	$39.18_{0.36}$	$34.45_{0.30}$	$29.07_{0.21}$	$10.55_{0.28}$	$7.63_{0.17}$	$6.52_{0.16}$
Radial SVM	$46.33_{0.47}$	$42.74_{0.42}$	$38.71_{0.43}$	$27.67_{1.15}$	$12.85_{0.90}$	$6.67_{0.20}$
Linear SVM	$42.38_{0.39}$	$40.96_{0.30}$	$39.55_{0.33}$	$19.41_{0.33}$	$17.05_{0.24}$	$15.49_{0.27}$
Radial GP	$40.60_{0.35}$	$36.79_{0.30}$	$32.21_{0.20}$	$22.53_{0.73}$	$17.78_{0.44}$	$14.43_{0.27}$
PenLDA	$44.37_{0.40}$	$43.47_{0.32}$	$42.50_{0.24}$	$21.20_{0.56}$	$19.83_{0.55}$	$19.81_{0.53}$
NSC	$44.74_{0.46}$	$43.59_{0.34}$	$42.37_{0.25}$	$22.62_{0.51}$	$19.11_{0.40}$	$17.52_{0.32}$
SCRDA	$44.09_{0.44}$	$42.02_{0.30}$	$40.08_{0.30}$	$19.71_{0.41}$	$16.74_{0.22}$	$15.28_{0.24}$
IR	$45.04_{0.39}$	$44.18_{0.33}$	$43.36_{0.26}$	$22.19_{0.58}$	$21.32_{0.53}$	$21.97_{0.55}$

Cardiac Arrhythmia diagnoses

The Cardiac Arrhythmia dataset[5] has one normal class of size 245, and 15 abnormal classes, which we combined to form the second class of size 206. We removed the nominal features and those with missing values, leaving $p = 194$ electrocardiogram (ECG) measurements. We see in the right panel of Table 2.7 that all random projection ensembles perform well, but for the larger sample size the random forest classifier is better.

2.7 Discussion and extensions

We have introduced a general framework for high-dimensional classification via the combination of the results of applying a base classifier on carefully selected low-dimensional random projections of the data. One of its attractive features is its generality: the approach can be used in conjunction with any base classifier. Moreover, although we explored in detail one method for combining the random projections (partly because it facilitates rigorous statistical analysis), there are many other options available here. For instance, instead of only retaining the projection within each block yielding the smallest estimate of test error, one might give weights to the different projections, where the weights decrease as the estimate of test error increases. Another interesting

[5] `https://archive.ics.uci.edu/ml/datasets/Arrhythmia`

Table 2.7: *Misclassification rates for the Musk and Cardiac Arrhythmia datasets.*

	Musk $p = 166$			Cardiac $p = 194$		
n	50	100	200	50	100	200
RP-LDA$_2$	$17.50_{0.36}$	$15.53_{0.25}$	$14.22_{0.18}$	$33.61_{0.36}$	$31.29_{0.31}$	$28.86_{0.26}$
RP-LDA$_5$	$18.81_{0.57}$	$14.86_{0.30}$	$12.09_{0.21}$	$33.18_{0.40}$	$29.88_{0.27}$	$27.47_{0.27}$
LDA	N/A	N/A	$24.88_{0.40}$	N/A	N/A	N/A
RP-QDA$_2$	$15.58_{0.32}$	$14.17_{0.26}$	$13.02_{0.18}$	$31.33_{0.30}$	$29.61_{0.24}$	$27.59_{0.26}$
RP-QDA$_5$	$14.70_{0.34}$	$12.72_{0.29}$	$9.93_{0.16}$	$\mathbf{30.84}_{0.30}$	$28.46_{0.24}$	$26.57_{0.23}$
QDA	N/A	N/A	N/A	N/A	N/A	N/A
RP-$k$nn$_2$	$14.92_{0.32}$	$12.33_{0.23}$	$10.09_{0.15}$	$32.78_{0.35}$	$30.58_{0.28}$	$28.08_{0.25}$
RP-$k$nn$_5$	$\mathbf{13.88}_{0.33}$	$\mathbf{10.96}_{0.30}$	$\mathbf{8.67}_{0.11}$	$33.40_{0.35}$	$30.73_{0.30}$	$27.24_{0.22}$
$k$nn	$16.22_{0.29}$	$14.41_{0.23}$	$11.14_{0.16}$	$40.63_{0.29}$	$38.94_{0.30}$	$35.76_{0.33}$
RF	$14.40_{0.17}$	$13.18_{0.16}$	$10.67_{0.16}$	$31.59_{0.35}$	$\mathbf{26.79}_{0.27}$	$\mathbf{22.61}_{0.27}$
Radial SVM	$15.27_{0.10}$	$15.25_{0.10}$	$15.21_{0.10}$	$48.37_{0.47}$	$47.23_{0.43}$	$46.85_{0.40}$
Linear SVM	$16.49_{0.35}$	$13.91_{0.22}$	$10.39_{0.15}$	$36.16_{0.45}$	$35.61_{0.36}$	$35.20_{0.32}$
Radial GP	$15.17_{0.11}$	$14.89_{0.12}$	$14.12_{0.16}$	$37.26_{0.39}$	$33.78_{0.36}$	$29.35_{0.31}$
PenLDA	$29.57_{0.72}$	$27.76_{0.22}$	$27.15_{0.53}$	N/A	N/A	N/A
NSC	$16.41_{0.34}$	$15.45_{0.15}$	$15.19_{0.10}$	$34.98_{0.44}$	$33.00_{0.37}$	$31.08_{0.38}$
SCRDA	$15.69_{0.34}$	$16.40_{0.52}$	$15.14_{0.22}$	$38.71_{0.44}$	$36.55_{0.45}$	$30.86_{0.39}$
IR	$32.22_{0.82}$	$30.83_{0.67}$	$30.58_{0.66}$	$32.05_{0.37}$	$30.29_{0.33}$	$28.67_{0.32}$

avenue to explore would be alternative methods for estimating the test error, such as sample splitting. The idea here would be to split the sample $\mathcal{T}_n$ into $\mathcal{T}_{n,1}$ and $\mathcal{T}_{n,2}$, say, where $|\mathcal{T}_{n,1}| = n^{(1)}$ and $|\mathcal{T}_{n,2}| = n^{(2)}$. We then use

$$\hat{L}^A_{n^{(1)},n^{(2)}} := \frac{1}{n^{(2)}} \sum_{(X_i,Y_i)\in\mathcal{T}_{n,2}} \mathbb{1}_{\{\hat{C}^A_{n^{(1)},\mathcal{T}^A_{n,1}}(X_i)\neq Y_i\}}$$

to estimate the test error $\mathcal{L}^A_{n^{(1)},1}$ based on the training data $\mathcal{T}_{n,1}$. Since $\mathcal{T}_{n,1}$ and $\mathcal{T}_{n,2}$ are independent, we can apply Hoeffding's inequality to deduce that

$$\sup_{A\in\mathcal{A}} \mathbb{P}\big\{|\mathcal{L}^A_{n^{(1)},1} - \hat{L}^A_{n^{(1)},n^{(2)}}| \geq \epsilon \;\big|\; \mathcal{T}_{n,1}\big\} \leq 2e^{-2n^{(2)}\epsilon^2}.$$

It then follows by very similar arguments to those given in Section 2.4.1 that

$$\begin{aligned}
\mathbb{E}\big(|\mathcal{L}^{A^*}_{n^{(1)}} - \hat{L}^{A^*}_{n^{(1)},n^{(2)}}| \;\big|\; \mathcal{T}_{n,1}\big) &\leq \Big(\frac{1+\log 2}{2n^{(2)}}\Big)^{1/2}, \\
\mathbb{E}\big(|\mathcal{L}^{A_1}_{n^{(1)}} - \hat{L}^{A_1}_{n^{(1)},n^{(2)}}| \;\big|\; \mathcal{T}_{n,1}\big) &\leq \Big(\frac{1+\log 2+\log B_2}{2n^{(2)}}\Big)^{1/2}.
\end{aligned} \tag{2.14}$$

The advantages of this approach are twofold: first, the bounds hold for any choice of base classifier (and still without any assumptions on the data generating mechanism); second, the bounds on the terms in (2.14) merely rely on Hoeffding's inequality as opposed to Vapnik–Chervonenkis theory, so are typically sharper. The disadvantage

is that the other terms in the bound in Theorem 2.5 will tend to be larger due to the reduced effective sample size. The choice of $n^{(1)}$ and $n^{(2)}$ in such an approach therefore becomes an interesting question.

Many practical classification problems involve $K > 2$ classes. The main issue in extending our methodology to such settings is the definition of $\hat{C}_n^{\mathrm{RP}}$ analogous to (2.2). To outline one approach, let

$$\hat{\nu}_{n,r}^{B_1}(x) := \frac{1}{B_1} \sum_{b_1=1}^{B_1} \mathbb{1}_{\{\hat{C}_n^{A_{b_1}}(x)=r\}}$$

for $r = 1, \ldots, K$. Given $\alpha_1, \ldots, \alpha_K > 0$ with $\sum_{r=1}^{K} \alpha_r = 1$, we can then define

$$\hat{C}_n^{\mathrm{RP}}(x) := \operatorname*{sargmax}_{r=1,\ldots,K} \{\alpha_r \hat{\nu}_{n,r}^{B_1}(x)\}.$$

The choice of $\alpha_1, \ldots, \alpha_K$ is analogous to the choice of α in the case $K = 2$. It is therefore natural to seek to minimise the the test error of the corresponding infinite-simulation random projection ensemble classifier as before.

In other situations, it may be advantageous to consider alternative types of projection, perhaps because of additional structure in the problem. One particularly interesting issue concerns ultrahigh-dimensional settings, say p in the thousands. Here, it may be too time-consuming to generate enough random projections to explore adequately the space $\mathcal{A}_{d\times p}$. As a mathematical quantification of this, the cardinality of an ϵ-net in the Euclidean norm of the surface of the Euclidean ball in $\mathbb{R}^p$ increases exponentially in p (e.g. Vershynin, 2012; Kim and Samworth, 2014). In such challenging problems, one might restrict the projections A to be axis-aligned, so that each row of A consists of a single non-zero component, equal to 1, and $p-1$ zero components. There are then only $\binom{p}{d} \leq p^d/d!$ choices for the projections, and if d is small, it may be feasible even to carry out an exhaustive search. Of course, this approach loses one of the attractive features of our original proposal, namely the fact that it is equivariant to orthogonal transformations. Nevertheless, corresponding theory can be obtained provided that the projection A^* in **(A.3)** is axis-aligned. This is a much stronger requirement, but it seems that imposing greater structure is inevitable to obtain good classification in such settings. A less restrictive option is to use sparse random projections: we first choose a subset of size $r < p$ of the variables and then use a Haar projection from the r-dimensional space to the d-dimensional space. The resulting method is again not equivariant to orthogonal transformations, but the corresponding projection A^* in **(A.3)** need only be sparse, not axis-aligned.

Finally here, we pose the question: are similar methods using random projection ensembles useful for other high-dimensional statistical problems, such as clustering or

regression?

2.8 Appendix

Proof of Theorem 2.1. Recall that the training data $\mathcal{T}_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ are fixed and the projections $A_1, A_2, \ldots,$ are independent and identically distributed in $\mathcal{A}$, independent of the pair (X, Y). The test error of the random projection ensemble classifier has the following representation:

$$\begin{aligned}\mathcal{L}(\hat{C}_n^{\mathrm{RP}}) &= \mathbb{P}\{\hat{C}_n^{\mathrm{RP}}(X) \neq Y\} = \pi_1 \mathbb{P}\{\hat{C}_n^{\mathrm{RP}}(X) = 2 | Y = 1\} + \pi_2 \mathbb{P}\{\hat{C}_n^{\mathrm{RP}}(X) = 1 | Y = 2\} \\ &= \pi_1 \mathbb{P}\big\{\hat{\nu}_n^{B_1}(X) < \alpha | Y = 1\big\} + \pi_2 \mathbb{P}\big\{\hat{\nu}_n^{B_1}(X) \geq \alpha | Y = 2\big\},\end{aligned}$$

where $\hat{\nu}_n^{B_1}(x)$ is defined in (2.1).

Let $U_{b_1} := \mathbb{1}_{\{\hat{C}_n^{A_{b_1}}(X)=1\}}$, for $b_1 = 1, \ldots B_1$. Then, conditional on $\hat{\mu}_n(X) = \theta \in [0, 1]$, the random variables $U_1, \ldots, U_{B_1}$ are independent, each having a Bernoulli(θ) distribution. Recall that $G_{n,1}$ and $G_{n,2}$ are the distribution functions of $\hat{\mu}_n(X)|\{Y = 1\}$ and $\hat{\mu}_n(X)|\{Y = 2\}$, respectively. We can therefore write

$$\begin{aligned}\mathbb{P}\big\{\hat{\nu}_n^{B_1}(X) < \alpha | Y = 1\big\} &= \int_0^1 \mathbb{P}\Big\{\frac{1}{B_1}\sum_{b_1=1}^{B_1} U_{b_1} < \alpha \Big| \hat{\mu}_n(X) = \theta\Big\}\, dG_{n,1}(\theta) \\ &= \int_0^1 \mathbb{P}(T < B_1\alpha)\, dG_{n,1}(\theta),\end{aligned}$$

where here and throughout the proof, T denotes a $\mathrm{Bin}(B_1, \theta)$ random variable. Similarly,

$$\mathbb{P}\big\{\hat{\nu}_n^{B_1}(X) \geq \alpha | Y = 2\big\} = 1 - \int_0^1 \mathbb{P}(T < B_1\alpha)\, dG_{n,2}(\theta).$$

It follows that

$$\mathcal{L}(\hat{C}_n^{\mathrm{RP}}) = \pi_2 + \int_0^1 \mathbb{P}(T < B_1\alpha)\, dG_n^\circ(\theta),$$

where $G_n^\circ := \pi_1 G_{n,1} - \pi_2 G_{n,2}$. Writing $g_n^\circ := \pi_1 g_{n,1} - \pi_2 g_{n,2}$, we now show that

$$\int_0^1 \big\{\mathbb{P}(T < B_1\alpha) - \mathbb{1}_{\{\theta<\alpha\}}\big\}\, dG_n^\circ(\theta) = \frac{1 - \alpha - [\![B_1\alpha]\!]}{B_1} g_n^\circ(\alpha) + \frac{\alpha(1-\alpha)}{2B_1}\dot{g}_n^\circ(\alpha) + o\Big(\frac{1}{B_1}\Big) \tag{2.15}$$

as $B_1 \to \infty$. Our proof involves a one-term Edgeworth expansion to the binomial distribution function in (2.15), where the error term is controlled uniformly in the parameter. The expansion relies on the following version of Esseen's smoothing lemma.

Theorem 2.6. *(Esseen, 1945, Chapter 2, Theorem 2b) Let c_1, C_1, $S > 0$, let $F : \mathbb{R} \to [0, \infty)$ be a non-decreasing function and let $G : \mathbb{R} \to \mathbb{R}$ be a function of bounded*

variation. Let $F^*(s) := \int_{-\infty}^{\infty} \exp(ist)\, dF(t)$ *and* $G^*(s) := \int_{-\infty}^{\infty} \exp(ist)\, dG(t)$ *be the Fourier–Stieltjes transforms of* F *and* G*, respectively. Suppose that*

- $\lim_{t\to-\infty} F(t) = \lim_{t\to-\infty} G(t) = 0$ *and* $\lim_{t\to\infty} F(t) = \lim_{t\to\infty} G(t)$*;*
- $\int_{-\infty}^{\infty} |F(t) - G(t)|\, dt < \infty$*;*
- *The set of discontinuities of* F *and* G *is contained in* $\{t_i : i \in \mathbb{Z}\}$*, where* (t_i) *is a strictly increasing sequence with* $\inf_i\{t_{i+1} - t_i\} \geq c_1$*; moreover* F *is constant on the intervals* $[t_i, t_{i+1})$ *for all* $i \in \mathbb{Z}$*;*
- $|\dot{G}(t)| \leq C_1$ *for all* $t \notin \{t_i : i \in \mathbb{Z}\}$.

Then there exist constants $c_2, C_2 > 0$ *such that*

$$\sup_{t\in\mathbb{R}} |F(t) - G(t)| \leq \frac{1}{\pi} \int_{-S}^{S} \left| \frac{F^*(s) - G^*(s)}{s} \right| ds + \frac{C_1 C_2}{S},$$

provided that $Sc_1 \geq c_2$.

Let $\sigma^2 := \theta(1-\theta)$, and let Φ and ϕ denote the standard normal distribution and density functions, respectively. Moreover, for $t \in \mathbb{R}$, let

$$p(t) = p(t, \theta) := \frac{(1-t^2)(1-2\theta)}{6\sigma},$$

and

$$q(t) = q(t, B_1, \theta) := \frac{1/2 - [\![B_1\theta + B_1^{1/2}\sigma t]\!]}{\sigma}.$$

In Proposition 2.7 below we apply Theorem 2.6 to the following functions:

$$F_{B_1}(t) = F_{B_1}(t, \theta) := \mathbb{P}\left(\frac{T - B_1\theta}{B_1^{1/2}\sigma} < t \right), \tag{2.16}$$

and

$$G_{B_1}(t) = G_{B_1}(t, \theta) := \Phi(t) + \phi(t) \frac{p(t, \theta) + q(t, B_1, \theta)}{B_1^{1/2}}. \tag{2.17}$$

Proposition 2.7. *Let* F_{B_1} *and* G_{B_1} *be as in* (2.16) *and* (2.17)*. There exists a constant* $C > 0$ *such that, for all* $B_1 \in \mathbb{N}$*,*

$$\sup_{\theta\in(0,1)} \sup_{t\in\mathbb{R}} \sigma^3 |F_{B_1}(t, \theta) - G_{B_1}(t, \theta)| \leq \frac{C}{B_1}.$$

Proposition 2.7, whose proof is given after the proof of Theorem 2.5, bounds uniformly in θ the error in the one-term Edgeworth expansion G_{B_1} of the distribution function F_{B_1}. Returning to the proof of Theorem 2.1, we will argue that the dominant contribution to the integral in (2.15) arises from the interval $(\max\{0, \alpha-\epsilon_1\}, \min\{\alpha+\epsilon_1, 1\})$,

where $\epsilon_1 := B_1^{-1/2} \log B_1$. For the remainder of the proof we assume B_1 is large enough that $[\alpha - \epsilon_1, \alpha + \epsilon_1] \subseteq (0, 1)$.

For the region $|\theta - \alpha| \geq \epsilon_1$, by Hoeffding's inequality, we have that

$$\begin{aligned}\sup_{|\theta-\alpha|\geq\epsilon_1} \left|\mathbb{P}(T < B_1\alpha) - \mathbb{1}_{\{\theta<\alpha\}}\right| &\leq \sup_{|\theta-\alpha|\geq\epsilon_1} \exp\bigl(-2B_1(\theta-\alpha)^2\bigr) \\ &\leq \exp(-2\log^2 B_1) = O(B_1^{-M}),\end{aligned}$$

for each $M > 0$, as $B_1 \to \infty$. It follows that

$$\begin{aligned}&\int_0^1 \bigl\{\mathbb{P}(T < B_1\alpha) - \mathbb{1}_{\{\theta<\alpha\}}\bigr\}\, dG_n^\circ(\theta) \\ &\qquad = \int_{\alpha-\epsilon_1}^{\alpha+\epsilon_1} \bigl\{\mathbb{P}(T < B_1\alpha) - \mathbb{1}_{\{\theta<\alpha\}}\bigr\}\, dG_n^\circ(\theta) + O(B_1^{-M}), \qquad (2.18)\end{aligned}$$

for each $M > 0$, as $B_1 \to \infty$.

For the region $|\theta - \alpha| < \epsilon_1$, by Proposition 2.7, there exists $C' > 0$ such that, for all B_1 sufficiently large,

$$\begin{aligned}\sup_{|\theta-\alpha|<\epsilon_1} \Biggl|\mathbb{P}(T < B_1\alpha) - \Phi\biggl(\frac{{B_1}^{1/2}(\alpha-\theta)}{\sigma}\biggr) \\ - \frac{1}{B_1^{1/2}}\phi\biggl(\frac{{B_1}^{1/2}(\alpha-\theta)}{\sigma}\biggr) r\biggl(\frac{{B_1}^{1/2}(\alpha-\theta)}{\sigma}\biggr)\Biggr| \leq \frac{C'}{B_1},\end{aligned}$$

where $r(t) := p(t) + q(t)$. Hence, using the fact that for large B_1, $\sup_{|\theta-\alpha|<\epsilon_1} |g_n^\circ(\theta)| \leq |g_n^\circ(\alpha)| + 1 < \infty$ under **(A.1)**, we have

$$\begin{aligned}&\int_{\alpha-\epsilon_1}^{\alpha+\epsilon_1} \bigl\{\mathbb{P}(T < B_1\alpha) - \mathbb{1}_{\{\theta<\alpha\}}\bigr\}\, dG_n^\circ(\theta) \\ &\qquad = \int_{\alpha-\epsilon_1}^{\alpha+\epsilon_1} \biggl\{\Phi\biggl(\frac{B_1^{1/2}(\alpha-\theta)}{\sigma}\biggr) - \mathbb{1}_{\{\theta<\alpha\}}\biggr\}\, dG_n^\circ(\theta) \\ &\qquad\quad + \frac{1}{B_1^{1/2}} \int_{\alpha-\epsilon_1}^{\alpha+\epsilon_1} \phi\biggl(\frac{B_1^{1/2}(\alpha-\theta)}{\sigma}\biggr) r\biggl(\frac{B_1^{1/2}(\alpha-\theta)}{\sigma}\biggr)\, dG_n^\circ(\theta) + o\Bigl(\frac{1}{B_1}\Bigr), \qquad (2.19)\end{aligned}$$

as $B_1 \to \infty$. To aid exposition, we will henceforth concentrate on the dominant terms in our expansions, denoting the remainder terms as $R_1, R_2, \ldots$. These remainders are

then controlled at the end of the argument. For the first term in (2.19), we write

$$\int_{\alpha-\epsilon_1}^{\alpha+\epsilon_1}\left\{\Phi\left(\frac{B_1^{1/2}(\alpha-\theta)}{\sigma}\right)-\mathbb{1}_{\{\theta<\alpha\}}\right\}dG_n^\circ(\theta)$$
$$=\int_{\alpha-\epsilon_1}^{\alpha+\epsilon_1}\left\{\Phi\left(\frac{B_1^{1/2}(\alpha-\theta)}{\sqrt{\alpha(1-\alpha)}}\right)-\mathbb{1}_{\{\theta<\alpha\}}\right\}dG_n^\circ(\theta)$$
$$+\frac{(1-2\alpha)B_1^{1/2}}{2\{\alpha(1-\alpha)\}^{3/2}}\int_{\alpha-\epsilon_1}^{\alpha+\epsilon_1}(\alpha-\theta)^2\phi\left(\frac{B_1^{1/2}(\alpha-\theta)}{\sqrt{\alpha(1-\alpha)}}\right)dG_n^\circ(\theta)+R_1. \tag{2.20}$$

Now, for the first term in (2.20),

$$\int_{\alpha-\epsilon_1}^{\alpha+\epsilon_1}\left\{\Phi\left(\frac{B_1^{1/2}(\alpha-\theta)}{\sqrt{\alpha(1-\alpha)}}\right)-\mathbb{1}_{\{\theta<\alpha\}}\right\}dG_n^\circ(\theta)$$
$$=\int_{\alpha-\epsilon_1}^{\alpha+\epsilon_1}\left\{\Phi\left(\frac{B_1^{1/2}(\alpha-\theta)}{\sqrt{\alpha(1-\alpha)}}\right)-\mathbb{1}_{\{\theta<\alpha\}}\right\}\left\{g_n^\circ(\alpha)+(\theta-\alpha)\dot{g}_n^\circ(\alpha)\right\}d\theta+R_2$$
$$=\frac{\sqrt{\alpha(1-\alpha)}}{B_1^{1/2}}\int_{-\infty}^{\infty}\{\Phi(-u)-\mathbb{1}_{\{u<0\}}\}\left\{g_n^\circ(\alpha)+\frac{\sqrt{\alpha(1-\alpha)}}{B_1^{1/2}}u\dot{g}_n^\circ(\alpha)\right\}du+R_2+R_3$$
$$=\frac{\alpha(1-\alpha)}{2B_1}\,\dot{g}_n^\circ(\alpha)+R_2+R_3. \tag{2.21}$$

For the second term in (2.20), write

$$\frac{(1-2\alpha)B_1^{1/2}}{2\{\alpha(1-\alpha)\}^{3/2}}\int_{\alpha-\epsilon_1}^{\alpha+\epsilon_1}(\alpha-\theta)^2\phi\left(\frac{B_1^{1/2}(\alpha-\theta)}{\sqrt{\alpha(1-\alpha)}}\right)dG_n^\circ(\theta)$$
$$=\frac{(1-2\alpha)B_1^{1/2}}{2\{\alpha(1-\alpha)\}^{3/2}}g_n^\circ(\alpha)\int_{\alpha-\epsilon_1}^{\alpha+\epsilon_1}(\alpha-\theta)^2\phi\left(\frac{B_1^{1/2}(\alpha-\theta)}{\sqrt{\alpha(1-\alpha)}}\right)d\theta+R_4$$
$$=\frac{1/2-\alpha}{B_1}g_n^\circ(\alpha)\int_{-\infty}^{\infty}u^2\phi(-u)\,du+R_4+R_5=\frac{1/2-\alpha}{B_1}g_n^\circ(\alpha)+R_4+R_5. \tag{2.22}$$

Returning to the second term in (2.19), observe that

$$\begin{aligned}
&\frac{1}{B_1^{1/2}} \int_{\alpha-\epsilon_1}^{\alpha+\epsilon_1} \phi\Big(\frac{B_1^{1/2}(\alpha-\theta)}{\sigma}\Big) r\Big(\frac{B_1^{1/2}(\alpha-\theta)}{\sigma}\Big)\, dG_n^\circ(\theta) \\
&\quad = \frac{1/2 - [\![B_1\alpha]\!]}{B_1^{1/2}} \int_{\alpha-\epsilon_1}^{\alpha+\epsilon_1} \frac{1}{\sigma}\phi\Big(\frac{B_1^{1/2}(\alpha-\theta)}{\sigma}\Big)\, dG_n^\circ(\theta) \\
&\qquad + \frac{1}{6B_1^{1/2}} \int_{\alpha-\epsilon_1}^{\alpha+\epsilon_1} \frac{(1-2\theta)}{\sigma}\Big\{1 - \frac{B_1(\alpha-\theta)^2}{\sigma^2}\Big\}\phi\Big(\frac{B_1^{1/2}(\alpha-\theta)}{\sigma}\Big)\, dG_n^\circ(\theta) \\
&\quad = \frac{1/2 - [\![B_1\alpha]\!]}{B_1^{1/2}} \int_{\alpha-\epsilon_1}^{\alpha+\epsilon_1} \frac{1}{\sigma}\phi\Big(\frac{B_1^{1/2}(\alpha-\theta)}{\sigma}\Big)\, dG_n^\circ(\theta) + R_6 \\
&\quad = \frac{1/2 - [\![B_1\alpha]\!]}{B_1^{1/2}\sqrt{\alpha(1-\alpha)}} g_n^\circ(\alpha) \int_{\alpha-\epsilon_1}^{\alpha+\epsilon_1} \phi\Big(\frac{B_1^{1/2}(\alpha-\theta)}{\sqrt{\alpha(1-\alpha)}}\Big)\, d\theta + R_6 + R_7 \\
&\quad = \frac{1/2 - [\![B_1\alpha]\!]}{B_1} g_n^\circ(\alpha) + R_6 + R_7 + R_8. \qquad (2.23)
\end{aligned}$$

The claim (2.15) will now follow from (2.18), (2.19), (2.20), (2.21), (2.22) and (2.23), once we have shown that

$$\sum_{j=1}^{8} |R_j| = o(B_1^{-1}) \qquad (2.24)$$

as $B_1 \to \infty$.

To bound R_1: For $\zeta \in (0,1)$, let $h_\theta(\zeta) := \Phi\big(\frac{B_1^{1/2}(\alpha-\theta)}{\sqrt{\zeta(1-\zeta)}}\big)$. Observe that, by a Taylor expansion about $\zeta = \alpha$, there exists $B_0 \in \mathbb{N}$, such that, for all $B_1 > B_0$ and all $\theta, \zeta \in (\alpha-\epsilon_1, \alpha+\epsilon_1)$,

$$\begin{aligned}
&\Big|\Phi\Big(\frac{B_1^{1/2}(\alpha-\theta)}{\sqrt{\zeta(1-\zeta)}}\Big) - \Phi\Big(\frac{B_1^{1/2}(\alpha-\theta)}{\sqrt{\alpha(1-\alpha)}}\Big) \\
&\qquad + (\zeta-\alpha)\frac{(1-2\alpha)B_1^{1/2}(\alpha-\theta)}{2\{\alpha(1-\alpha)\}^{3/2}}\phi\Big(\frac{B_1^{1/2}(\alpha-\theta)}{\sqrt{\alpha(1-\alpha)}}\Big)\Big| \\
&\qquad\quad = |h_\theta(\zeta) - h_\theta(\alpha) - (\zeta-\alpha)\dot{h}_\theta(\alpha)| \\
&\qquad\quad \le \frac{(\zeta-\alpha)^2}{2} \sup_{\zeta' \in [\alpha-\zeta, \alpha+\zeta]} |\ddot{h}_\theta(\zeta')| \le (\zeta-\alpha)^2 \frac{\log^3 B_1}{2\sqrt{2\pi}\{\alpha(1-\alpha)\}^{7/2}}.
\end{aligned}$$

Using this bound with $\zeta = \theta$, we deduce that, for all B_1 sufficiently large,

$$\begin{aligned}
|R_1| &= \left| \int_{\alpha-\epsilon_1}^{\alpha+\epsilon_1} \left\{ \Phi\Big(\frac{B_1^{1/2}(\alpha-\theta)}{\sigma} \Big) - \Phi\Big(\frac{B_1^{1/2}(\alpha-\theta)}{\sqrt{\alpha(1-\alpha)}} \Big) \right. \right. \\
&\qquad \left. \left. - \frac{(1-2\alpha)B_1^{1/2}(\alpha-\theta)^2}{2\{\alpha(1-\alpha)\}^{3/2}} \phi\Big(\frac{B_1^{1/2}(\alpha-\theta)}{\sqrt{\alpha(1-\alpha)}} \Big) \right\} dG_n^\circ(\theta) \right| \\
&\leq \frac{\log^3 B_1}{2\sqrt{2\pi}\{\alpha(1-\alpha)\}^{7/2}} \int_{\alpha-\epsilon_1}^{\alpha+\epsilon_1} (\theta-\alpha)^2 |g_n^\circ(\theta)| \, d\theta \\
&\leq \frac{\log^6 B_1}{3\sqrt{2\pi}B_1^{3/2}\{\alpha(1-\alpha)\}^{7/2}} \sup_{|\theta-\alpha|\leq\epsilon_1} |g_n^\circ(\theta)| = o\Big(\frac{1}{B_1}\Big)
\end{aligned}$$

as $B_1 \to \infty$.

To bound R_2: Since g_n° is differentiable at α, given $\epsilon > 0$, there exists $\delta_\epsilon > 0$ such that

$$|g_n^\circ(\theta) - g_n^\circ(\alpha) - (\theta-\alpha)\dot{g}_n^\circ(\alpha)| < \epsilon|\theta-\alpha|,$$

for all $|\theta-\alpha| < \delta_\epsilon$. It follows that, for all B_1 sufficiently large,

$$\begin{aligned}
|R_2| &= \left| \int_{\alpha-\epsilon_1}^{\alpha+\epsilon_1} \left\{ \Phi\Big(\frac{B_1^{1/2}(\alpha-\theta)}{\sqrt{\alpha(1-\alpha)}} \Big) - \mathbb{1}_{\{\theta<\alpha\}} \right\} dG_n^\circ(\theta) \right. \\
&\qquad \left. - \int_{\alpha-\epsilon_1}^{\alpha+\epsilon_1} \left\{ \Phi\Big(\frac{B_1^{1/2}(\alpha-\theta)}{\sqrt{\alpha(1-\alpha)}} \Big) - \mathbb{1}_{\{\theta<\alpha\}} \right\} \{ g_n^\circ(\alpha) + (\theta-\alpha)\dot{g}_n^\circ(\alpha) \} \, d\theta \right| \\
&\leq \epsilon \int_{\alpha-\epsilon_1}^{\alpha+\epsilon_1} \left| \Phi\Big(\frac{B_1^{1/2}(\alpha-\theta)}{\sqrt{\alpha(1-\alpha)}} \Big) - \mathbb{1}_{\{\theta<\alpha\}} \right| |\theta-\alpha| \, d\theta \\
&\leq \frac{\epsilon\alpha(1-\alpha)}{B_1} \int_{-\log B_1/\sqrt{\alpha(1-\alpha)}}^{\log B_1/\sqrt{\alpha(1-\alpha)}} \big| \Phi(-u) - \mathbb{1}_{\{u<0\}} \big| |u| \, du \\
&\leq \frac{2\epsilon\alpha(1-\alpha)}{B_1} \int_0^\infty u\Phi(-u) \, du = \frac{\epsilon\alpha(1-\alpha)}{2B_1}.
\end{aligned}$$

We deduce that $|R_2| = o(B_1^{-1})$ as $B_1 \to \infty$.

To bound R_3: For large B_1, we have

$$\begin{aligned}
|R_3| &= \left| \int_{\alpha-\epsilon_1}^{\alpha+\epsilon_1} \left\{ \Phi\Big(\frac{B_1^{1/2}(\alpha-\theta)}{\sqrt{\alpha(1-\alpha)}} \Big) - \mathbb{1}_{\{\theta<\alpha\}} \right\} \{ g_n^\circ(\alpha) + (\theta-\alpha)\dot{g}_n^\circ(\alpha) \} \, d\theta \right. \\
&\qquad \left. - \frac{\sqrt{\alpha(1-\alpha)}}{B_1^{1/2}} \int_{-\infty}^\infty \{ \Phi(-u) - \mathbb{1}_{\{u<0\}} \} \left\{ g_n^\circ(\alpha) + \frac{\sqrt{\alpha(1-\alpha)}}{B_1^{1/2}} u\dot{g}_n^\circ(\alpha) \right\} du \right| \\
&= \frac{2\alpha(1-\alpha)}{B_1} |\dot{g}_n^\circ(\alpha)| \int_{\epsilon_1 B_1^{1/2}/\{\alpha(1-\alpha)\}^{1/2}}^\infty u\Phi(-u) \, du \\
&\leq \frac{2\{\alpha(1-\alpha)\}^{3/2}}{B_1 \log B_1} |\dot{g}_n^\circ(\alpha)| \int_0^\infty u^2\Phi(-u) \, du = \frac{2\sqrt{2}\{\alpha(1-\alpha)\}^{3/2}}{3\sqrt{\pi}B_1 \log B_1} |\dot{g}_n^\circ(\alpha)| = o(B_1^{-1})
\end{aligned}$$

as $B_1 \to \infty$.

To bound R_4: By the bound in (2.25), we have that, given $\epsilon > 0$, for all B_1 sufficiently large,

$$\begin{aligned}|R_4| &= \left| \frac{(1-2\alpha)B_1^{1/2}}{2\{\alpha(1-\alpha)\}^{3/2}} \int_{\alpha-\epsilon_1}^{\alpha+\epsilon_1} (\alpha-\theta)^2 \phi\Big(\frac{B_1^{1/2}(\alpha-\theta)}{\sqrt{\alpha(1-\alpha)}} \Big) \{g_n^\circ(\theta) - g_n^\circ(\alpha)\}\, d\theta \right| \\ &\leq \frac{\epsilon|1-2\alpha|}{2B_1} \int_{-\infty}^{\infty} u^2 \phi(-u)\, du = \frac{\epsilon|1-2\alpha|}{2B_1}.\end{aligned}$$

To bound R_5: For all B_1 sufficiently large,

$$\begin{aligned}|R_5| &= \frac{|1-2\alpha|}{B_1} |g_n^\circ(\alpha)| \int_{\log B_1/\sqrt{\alpha(1-\alpha)}}^{\infty} u^2 \phi(-u)\, du \\ &\leq \frac{\sqrt{\alpha(1-\alpha)}}{B_1 \log B_1} |g_n^\circ(\alpha)| \int_0^{\infty} u^3 \phi(-u)\, du = \frac{\sqrt{2\alpha(1-\alpha)}}{\sqrt{\pi} B_1 \log B_1} |g_n^\circ(\alpha)| = o\Big(\frac{1}{B_1}\Big)\end{aligned}$$

as $B_1 \to \infty$.

To bound R_6: We write $R_6 = R_{61} + R_{62}$, where

$$R_{61} := \frac{(1-2\alpha)}{6B_1^{1/2}\sqrt{\alpha(1-\alpha)}} \int_{\alpha-\epsilon_1}^{\alpha+\epsilon_1} \left\{ 1 - \frac{B_1(\alpha-\theta)^2}{\alpha(1-\alpha)} \right\} \phi\Big(\frac{B_1^{1/2}(\alpha-\theta)}{\sqrt{\alpha(1-\alpha)}} \Big)\, dG_n^\circ(\theta)$$

and

$$\begin{aligned}R_{62} := &\frac{1}{6B_1^{1/2}} \int_{\alpha-\epsilon_1}^{\alpha+\epsilon_1} \frac{(1-2\theta)}{\sigma} \left\{ 1 - \frac{B_1(\alpha-\theta)^2}{\sigma^2} \right\} \phi\Big(\frac{B_1^{1/2}(\alpha-\theta)}{\sigma} \Big)\, dG_n^\circ(\theta) \\ &- \frac{(1-2\alpha)}{6B_1^{1/2}\sqrt{\alpha(1-\alpha)}} \int_{\alpha-\epsilon_1}^{\alpha+\epsilon_1} \left\{ 1 - \frac{B_1(\alpha-\theta)^2}{\alpha(1-\alpha)} \right\} \phi\Big(\frac{B_1^{1/2}(\alpha-\theta)}{\sqrt{\alpha(1-\alpha)}} \Big)\, dG_n^\circ(\theta).\end{aligned}$$

Since g_n° is continuous at α, given $\epsilon > 0$, there exists $B_0' \in \mathbb{N}$ such that, for all $B_1 > B_0'$,

$$\sup_{|\theta-\alpha| \leq \epsilon_1} |g_n^\circ(\theta) - g_n^\circ(\alpha)| < \epsilon. \tag{2.25}$$

It follows that, for $B_1 > B_0'$,

$$\begin{aligned}|R_{61}| &\le \frac{|1-2\alpha|}{6B_1^{1/2}\sqrt{\alpha(1-\alpha)}}|g_n^\circ(\alpha)|\left|\int_{\alpha-\epsilon_1}^{\alpha+\epsilon_1}\left\{1-\frac{B_1(\alpha-\theta)^2}{\alpha(1-\alpha)}\right\}\phi\Big(\frac{B_1^{1/2}(\alpha-\theta)}{\sqrt{\alpha(1-\alpha)}}\Big)\,d\theta\right| \\ &\quad + \epsilon\frac{|1-2\alpha|}{6B_1^{1/2}\sqrt{\alpha(1-\alpha)}}\int_{\alpha-\epsilon_1}^{\alpha+\epsilon_1}\left|1-\frac{B_1(\alpha-\theta)^2}{\alpha(1-\alpha)}\right|\phi\Big(\frac{B_1^{1/2}(\alpha-\theta)}{\sqrt{\alpha(1-\alpha)}}\Big)\,d\theta. \\ &\le \frac{|1-2\alpha|}{6B_1}|g_n^\circ(\alpha)|\left|\int_{-\log B_1/\sqrt{\alpha(1-\alpha)}}^{\log B_1/\sqrt{\alpha(1-\alpha)}}(1-u^2)\phi(-u)\,du\right| \\ &\quad + \epsilon\frac{|1-2\alpha|}{6B_1}\int_{-\infty}^{\infty}(1+u^2)\phi(-u)\,du \le \frac{\epsilon}{B_1}\end{aligned}$$

for all sufficiently large B_1. We deduce that $R_{61} = o(B_1^{-1})$ as $B_1 \to \infty$.

To control R_{62}, by the mean value theorem, we have that for all B_1 sufficiently large and all $\zeta \in [\alpha-\epsilon_1, \alpha+\epsilon_1]$,

$$\begin{aligned}\sup_{|\theta-\alpha|<\epsilon_1}\Bigg|\frac{(1-2\zeta)}{\sqrt{\zeta(1-\zeta)}}\left\{1-\frac{B_1(\alpha-\theta)^2}{\zeta(1-\zeta)}\right\}\phi\Big(\frac{B_1^{1/2}(\alpha-\theta)}{\sqrt{\zeta(1-\zeta)}}\Big) \\ - \frac{(1-2\alpha)}{\sqrt{\alpha(1-\alpha)}}\left\{1-\frac{B_1(\alpha-\theta)^2}{\alpha(1-\alpha)}\right\}\phi\Big(\frac{B_1^{1/2}(\alpha-\theta)}{\sqrt{\alpha(1-\alpha)}}\Big)\Bigg| \\ \le \frac{\log^4 B_1}{\sqrt{2\pi}\{\alpha(1-\alpha)\}^{7/2}}|\zeta-\alpha|.\end{aligned}$$

Thus, for large B_1,

$$\begin{aligned}|R_{62}| &\le \frac{\log^4 B_1}{6\sqrt{2\pi}B_1^{1/2}\{\alpha(1-\alpha)\}^{7/2}}\sup_{|\theta-\alpha|\le\epsilon_1}|g_n^\circ(\theta)|\int_{\alpha-\epsilon_1}^{\alpha+\epsilon_1}|\theta-\alpha|\,d\theta \\ &\le \frac{\log^6 B_1\{1+|g_n^\circ(\alpha)|\}}{6\sqrt{2\pi}B_1^{3/2}\{\alpha(1-\alpha)\}^{7/2}} = o\Big(\frac{1}{B_1}\Big).\end{aligned}$$

We deduce that $|R_6| = o(B_1^{-1})$ as $B_1 \to \infty$.

To bound R_7: write $R_7 = R_{71} + R_{72}$, where

$$R_{71} := \frac{1/2 - [\![B_1\alpha]\!]}{B_1^{1/2}\sqrt{\alpha(1-\alpha)}}\int_{\alpha-\epsilon_1}^{\alpha+\epsilon_1}\phi\Big(\frac{B_1^{1/2}(\alpha-\theta)}{\sqrt{\alpha(1-\alpha)}}\Big)\{g_n^\circ(\theta)-g_n^\circ(\alpha)\}\,d\theta,$$

and

$$\begin{aligned}R_{72} := \frac{1/2 - [\![B_1\alpha]\!]}{B_1^{1/2}}\int_{\alpha-\epsilon_1}^{\alpha+\epsilon_1}\Bigg\{\frac{1}{\sigma}\phi\Big(\frac{B_1^{1/2}(\alpha-\theta)}{\sigma}\Big) \\ - \frac{1}{\sqrt{\alpha(1-\alpha)}}\phi\Big(\frac{B_1^{1/2}(\alpha-\theta)}{\sqrt{\alpha(1-\alpha)}}\Big)\Bigg\}\,dG_n^\circ(\theta).\end{aligned}$$

By the bound in (2.25), given $\epsilon > 0$, for all B_1 sufficiently large,

$$|R_{71}| \leq \frac{\epsilon}{2B_1^{1/2}\sqrt{\alpha(1-\alpha)}} \int_{-\infty}^{\infty} \phi\Big(\frac{B_1^{1/2}(\alpha-\theta)}{\sqrt{\alpha(1-\alpha)}}\Big)\, d\theta = \frac{\epsilon}{2B_1}.$$

Moreover, by the mean value theorem, for all B_1 sufficiently large and all $|\zeta - \alpha| \leq \epsilon_1$,

$$\sup_{|\theta-\alpha|<\epsilon_1} \left| \frac{1}{\sqrt{\zeta(1-\zeta)}}\phi\Big(\frac{B_1^{1/2}(\alpha-\theta)}{\sqrt{\zeta(1-\zeta)}}\Big) - \frac{1}{\sqrt{\alpha(1-\alpha)}}\phi\Big(\frac{B_1^{1/2}(\alpha-\theta)}{\sqrt{\alpha(1-\alpha)}}\Big)\right| \leq \frac{\log^2 B_1}{\sqrt{2\pi}\{\alpha(1-\alpha)\}^{5/2}}|\zeta-\alpha|.$$

It follows that, for all B_1 sufficiently large,

$$\begin{aligned} |R_{72}| &\leq \frac{\log^2 B_1}{2\sqrt{2\pi}B_1^{1/2}\{\alpha(1-\alpha)\}^{5/2}} \sup_{|\theta-\alpha|\leq\epsilon_1} |g_n^\circ(\theta)| \int_{\alpha-\epsilon_1}^{\alpha+\epsilon_1} |\theta-\alpha|\, d\theta \\ &\leq \frac{\log^4 B_1\{1+|g_n^\circ(\alpha)|\}}{2\sqrt{2\pi}B_1^{3/2}\{\alpha(1-\alpha)\}^{5/2}}. \end{aligned}$$

We deduce that $|R_7| = o(B_1^{-1})$ as $B_1 \to \infty$.

To bound R_8: We have

$$|R_8| = \frac{2(1/2 - [\![B_1\alpha]\!])}{B_1}|g_n^\circ(\alpha)| \int_{\epsilon_1 B_1^{1/2}/\{\alpha(1-\alpha)\}^{1/2}}^{\infty} \phi(-u)\, du = o\Big(\frac{1}{B_1}\Big)$$

as $B_1 \to \infty$.

We have now established the claim at (2.24), and the result follows. □

Proof of Theorem 2.2. We have

$$\begin{aligned} \mathcal{L}(\hat{C}_n^{\mathrm{RP}^*}) - R(C^{\mathrm{Bayes}}) &= \mathbb{E}[\mathbb{P}\{\hat{C}_n^{\mathrm{RP}^*}(X) \neq Y|X\} - \mathbb{P}\{C^{\mathrm{Bayes}}(X) \neq Y|X\}] \\ &= \mathbb{E}\big[\eta(X)(\mathbb{1}_{\{\hat{C}_n^{\mathrm{RP}^*}(X)=2\}} - \mathbb{1}_{\{C^{\mathrm{Bayes}}(X)=2\}}) \\ &\qquad + \{1-\eta(X)\}(\mathbb{1}_{\{\hat{C}_n^{\mathrm{RP}^*}(X)=1\}} - \mathbb{1}_{\{C^{\mathrm{Bayes}}(X)=1\}})\big] \\ &= \mathbb{E}\big\{|2\eta(X)-1||\mathbb{1}_{\{\hat{\mu}_n^{B_2}(X)<\alpha\}} - \mathbb{1}_{\{\eta(X)<1/2\}}|\big\} \\ &= \mathbb{E}\big\{|2\eta(X)-1|\mathbb{1}_{\{\hat{\mu}_n^{B_2}(X)\geq\alpha\}}\mathbb{1}_{\{\eta(X)<1/2\}}\big\} \\ &\qquad + \mathbb{E}\big\{|2\eta(X)-1|\mathbb{1}_{\{\hat{\mu}_n^{B_2}(X)<\alpha\}}\mathbb{1}_{\{\eta(X)\geq1/2\}}\big\} \\ &\leq \frac{1}{\alpha}\mathbb{E}\big\{|2\eta(X)-1|\hat{\mu}_n^{B_2}(X)\mathbb{1}_{\{\eta(X)<1/2\}}\big\} \\ &\qquad + \frac{1}{1-\alpha}\mathbb{E}\big[|2\eta(X)-1|\{1-\hat{\mu}_n^{B_2}(X)\}\mathbb{1}_{\{\eta(X)\geq1/2\}}\big] \\ &\leq \frac{1}{\min(\alpha,1-\alpha)}\mathbb{E}\big\{|2\eta(X)-1||1-\hat{\mu}_n^{B_2}(X) - \mathbb{1}_{\{\eta(X)<1/2\}}|\big\}. \end{aligned}$$

Now, for each $x \in \mathbb{R}^p$,

$$\begin{aligned}|1 - \hat{\mu}_n^{B_2}(x) - \mathbb{1}_{\{\eta(x)<1/2\}}| &= |\mathbb{P}\{\hat{C}_n^{A_1}(x) = 2\} - \mathbb{1}_{\{\eta(x)<1/2\}}| \\ &= \mathbb{E}|\mathbb{1}_{\{\hat{C}_n^{A_1}(x)=2\}} - \mathbb{1}_{\{\eta(x)<1/2\}}|.\end{aligned}$$

We deduce that

$$\begin{aligned}&\mathbb{E}\big\{|2\eta(X) - 1||1 - \hat{\mu}_n^{B_2}(X) - \mathbb{1}_{\{\eta(X)<1/2\}}|\big\} \\ &\qquad = \mathbb{E}\Big[\mathbb{E}\big\{|2\eta(X) - 1||\mathbb{1}_{\{\hat{C}_n^{A_1}(X)=2\}} - \mathbb{1}_{\{\eta(X)<1/2\}}|\big|X\big\}\Big] \\ &\qquad = \mathbb{E}\Big[\mathbb{E}\big\{|2\eta(X) - 1||\mathbb{1}_{\{\hat{C}_n^{A_1}(X)=2\}} - \mathbb{1}_{\{\eta(X)<1/2\}}|\big|A_1\big\}\Big] \\ &\qquad = \mathbb{E}(\mathcal{L}_n^{A_1}) - R(C^{\text{Bayes}}).\end{aligned}$$

The result follows. □

Proof of Proposition 2.3. First write

$$\mathbb{E}(\mathcal{L}_n^{A_1}) - R(C^{\text{Bayes}}) = \mathbb{E}(\hat{L}_n^{A_1}) - R(C^{\text{Bayes}}) + \epsilon_n.$$

Using (**A.2**), we have that

$$\begin{aligned}\mathbb{E}(\hat{L}_n^{A_1}) &= \hat{L}_n^* + \frac{1}{n}\sum_{j=0}^{\lfloor n(1-\hat{L}_n^*)\rfloor - 1}\{1 - \beta_n(j)\}^{B_2} \\ &\leq \hat{L}_n^* + \frac{1}{n}\sum_{j=0}^{J}\Big(1 - \beta_0 - \beta\frac{j^\rho}{n^\rho}\Big)^{B_2} + \frac{1}{n}\sum_{j=J+1}^{\lfloor n(1-\hat{L}_n^*)\rfloor - 1}\{1 - \beta_n(j)\}^{B_2},\end{aligned}$$

where $J := \big\lfloor n\big(\frac{\log^2 B_2}{\beta B_2}\big)^{1/\rho}\big\rfloor$. Now,

$$\begin{aligned}\frac{1}{n}\sum_{j=0}^{J}\Big(1 - \beta_0 - \beta\frac{j^\rho}{n^\rho}\Big)^{B_2} &\leq \frac{(1-\beta_0)^{B_2}}{n} + \int_0^{J/n}(1 - \beta_0 - \beta x^\rho)^{B_2}\,dx \\ &\leq (1-\beta_0)^{B_2}\bigg\{\frac{1}{n} + \int_0^{J/n}\exp\Big(-\frac{B_2\beta x^\rho}{1-\beta_0}\Big)\,dx\bigg\} \\ &\leq (1-\beta_0)^{B_2}\bigg\{\frac{1}{n} + \frac{(1-\beta_0)^{1/\rho}\Gamma(1+1/\rho)}{B_2^{1/\rho}\beta^{1/\rho}}\bigg\}.\end{aligned}$$

Moreover, since $(J+1)^\rho/n^\rho \geq \log^2 B_2/(\beta B_2)$, we have

$$\frac{1}{n}\sum_{j=J+1}^{\lfloor n(1-\hat{L}_n^*)\rfloor - 1}\{1 - \beta_n(j)\}^{B_2} \leq \Big(1 - \beta_0 - \frac{\log^2 B_2}{B_2}\Big)^{B_2} \leq (1-\beta_0)^{B_2}\exp\Big(-\frac{\log^2 B_2}{1-\beta_0}\Big).$$

The result follows. □

Proof of Proposition 2.4. For a Borel set $C \subseteq \mathbb{R}^d$, let $P_{A^*X}(C) := \int_{\{x:A^*x\in C\}} dP_X(x)$, so that P_{A^*X} is the marginal distribution of A^*X. Further, for $z \in \mathbb{R}^d$, write $P_{X|A^*X=z}$ for the conditional distribution of X given $A^*X = z$. If Y is independent of X given A^*X, and if B is a Borel subset of $\mathbb{R}^p$, then

$$\begin{aligned}\int_B \eta^{A^*}(A^*x)\,dP_X(x) &= \int_{\mathbb{R}^d}\int_{B\cap\{x:A^*x=z\}} \eta^{A^*}(A^*x)\,dP_{X|A^*X=z}(w)\,dP_{A^*X}(z)\\ &= \int_{\mathbb{R}^d} \eta^{A^*}(z)\mathbb{P}(X\in B|A^*X=z)\,dP_{A^*X}(z)\\ &= \int_{\mathbb{R}^d} \mathbb{P}(Y=1, X\in B|A^*X=z)\,dP_{A^*X}(z)\\ &= \mathbb{P}(Y=1, X\in B) = \int_B \eta(x)\,dP_X(x).\end{aligned}$$

We deduce that $P_X(\{x \in \mathbb{R}^p : \eta(x) \neq \eta^{A^*}(A^*x)\}) = 0$; in particular, **(A.3)** holds, as required. □

Proof of Theorem 2.5. By the definitions of $\hat{L}_n^*$ and $\epsilon_n^{A^*}$, we have $\hat{L}_n^* \leq \hat{L}_n^{A^*} = \mathcal{L}_n^{A^*} - \epsilon_n^{A^*}$. Moreover,

$$\begin{aligned}R^{A^*-\mathrm{Bayes}} &= \int_{\mathbb{R}^p\times\{1,2\}} \mathbb{1}_{\{C^{A^*-\mathrm{Bayes}}(A^*x)\neq y\}}\,dP(x,y)\\ &= \int_{\mathbb{R}^p} \eta(x)\mathbb{1}_{\{\eta^{A^*}(A^*x)<1/2\}}\,dP_X(x) + \int_{\mathbb{R}^p}\{1-\eta(x)\}\mathbb{1}_{\{\eta^{A^*}(A^*x)\geq 1/2\}}\,dP_X(x)\\ &= \int_{\mathbb{R}^p} \eta(x)\mathbb{1}_{\{\eta(x)<1/2\}}\,dP_X(x) + \int_{\mathbb{R}^p}\{1-\eta(x)\}\mathbb{1}_{\{\eta(x)\geq 1/2\}}\,dP_X(x)\\ &= R(C^{\mathrm{Bayes}}).\end{aligned}$$

Note that we have used **(A.3)** to obtain the penultimate equality. The result now follows immediately from these facts, together with Theorem 2.1, Theorem 2.2 and Proposition 2.3. □

Proof of Proposition 2.7. Recall that $\sigma^2 := \theta(1-\theta)$. Let

$$\begin{aligned}F^*_{B_1}(s) = F^*_{B_1}(s,\theta) &:= \int_{-\infty}^{\infty} e^{ist}\,dF_{B_1}(t)\\ &= \left\{(1-\theta)\exp\left(-\frac{is\theta}{B_1^{1/2}\sigma}\right) + \theta\exp\left(\frac{is(1-\theta)}{B_1^{1/2}\sigma}\right)\right\}^{B_1}.\end{aligned}$$

Moreover, let $P(t) := \frac{\phi(t)p(t)}{B_1^{1/2}}$ and $Q(t) := \frac{\phi(t)q(t)}{B_1^{1/2}}$. By, for example, Gnedenko and

Kolmogorov (1954, Chapter 8, Section 43), we have

$$\Phi^*(s) := \int_{-\infty}^{\infty} \exp(ist)\, d\Phi(t) = \exp(-s^2/2),$$

$$P^*(s) := \int_{-\infty}^{\infty} \exp(ist)\, dP(t) = -\frac{1-2\theta}{6B_1^{1/2}\sigma} is^3 \exp(-s^2/2)$$

and

$$\begin{aligned} Q^*(s) &:= \int_{-\infty}^{\infty} \exp(ist)\, dQ(t) \\ &= -\frac{s}{2\pi B_1^{1/2}\sigma} \sum_{l\in\mathbb{Z}\setminus\{0\}} \frac{\exp(i2\pi B_1 l\theta)}{l} \exp\Big\{-\frac{1}{2}\big(s + 2\pi B_1^{1/2}\sigma l\big)^2\Big\}. \end{aligned}$$

Thus

$$\begin{aligned} G^*_{B_1}(s) = G^*_{B_1}(s,\theta) &:= \int_{-\infty}^{\infty} \exp(ist)\, dG_{B_1}(t) = \Phi^*(s) + P^*(s) + Q^*(s) \\ &= \exp(-s^2/2) - \frac{1-2\theta}{6B_1^{1/2}\sigma} is^3 \exp(-s^2/2) \\ &\quad - \frac{s}{2\pi B_1^{1/2}\sigma} \sum_{l\in\mathbb{Z}\setminus\{0\}} \frac{\exp(i2\pi B_1 l\theta)}{l} \exp\Big\{-\frac{1}{2}\big(s + 2\pi B_1^{1/2}\sigma l\big)^2\Big\}. \end{aligned}$$

Letting $c_2 > 0$ be the constant given in the statement of Theorem 2.6 (in fact we assume without loss of generality that $c_2 > \pi$), we show that there exists a constant $C' > 0$ such that, for all $B_1 \in \mathbb{N}$,

$$\sup_{\theta\in(0,1)} \sigma^3 \int_{-c_2 B_1^{1/2}\sigma}^{c_2 B_1^{1/2}\sigma} \left| \frac{F^*_{B_1}(s,\theta) - G^*_{B_1}(s,\theta)}{s} \right| ds \le \frac{C'}{B_1}. \tag{2.26}$$

To show (2.26), write

$$\begin{aligned} \int_{-c_2 B_1^{1/2}\sigma}^{c_2 B_1^{1/2}\sigma} \left| \frac{F^*_{B_1}(s) - G^*_{B_1}(s)}{s} \right| ds &= \int_{-S_1}^{S_1} \left| \frac{F^*_{B_1}(s) - G^*_{B_1}(s)}{s} \right| ds \\ &\quad + \int_{S_1\le|s|\le S_2} \left| \frac{F^*_{B_1}(s) - G^*_{B_1}(s)}{s} \right| ds + \int_{S_2\le|s|\le c_2 B_1^{1/2}\sigma} \left| \frac{F^*_{B_1}(s) - G^*_{B_1}(s)}{s} \right| ds, \end{aligned} \tag{2.27}$$

where $S_1 := \frac{B_1^{1/2}\sigma^{3/2}}{32(3\theta^2-3\theta+1)^{3/4}}$ and $S_2 := \pi B_1^{1/2}\sigma$. Note that $S_1 \le S_2/2$ for all $\theta \in (0,1)$.

We bound each term in (2.27) in turn. By Gnedenko and Kolmogorov (1954, Theorem 1, Section 41), there exists a universal constant $C_3 > 0$, such that, for all

$|s| \leq S_1$,

$$|F^*_{B_1}(s,\theta) - \Phi^*(s) - P^*(s)| \leq \frac{C_3}{B_1\sigma^3}(s^4 + s^6)\exp(-s^2/4).$$

Thus

$$\int_{-S_1}^{S_1}\left|\frac{F^*_{B_1}(s) - \Phi^*(s) - P^*(s)}{s}\right| ds \leq \frac{C_3}{B_1\sigma^3}\int_{-\infty}^{\infty}(|s|^3 + |s|^5)\exp(-s^2/4)\,ds = \frac{144C_3}{B_1\sigma^3}. \tag{2.28}$$

Moreover, observe that $\left(s + 2\pi B_1^{1/2}\sigma l\right)^2 \geq s^2 + 2\pi^2 B_1\sigma^2 l^2$ for all $|s| \leq S_1$. Thus, for $|s| \leq S_1$,

$$\begin{aligned}\left|\frac{Q^*(s)}{s}\right| &\leq \frac{1}{2\pi B_1^{1/2}\sigma}\left|\sum_{l\in\mathbb{Z}\setminus\{0\}}\frac{\exp(i2\pi B_1 l\theta)}{l}\exp\Big\{-\frac{1}{2}\big(s + 2\pi B_1^{1/2}\sigma l\big)^2\Big\}\right| \\ &\leq \frac{\phi(s)}{\sqrt{2\pi}B_1^{1/2}\sigma}\int_{-\infty}^{\infty}\exp\big(-\pi^2 B_1\sigma^2 u^2\big)\,du = \frac{\phi(s)}{\sqrt{2\pi}B_1\sigma^2}.\end{aligned}$$

It follows that

$$\int_{-S_1}^{S_1}\left|\frac{Q^*(s)}{s}\right| ds \leq \frac{1}{\sqrt{2}\pi B_1\sigma^2}. \tag{2.29}$$

For $|s| \in [S_1, S_2]$, observe that

$$|F^*_{B_1}(s)| = \left[1 - 2\sigma^2\left\{1 - \cos\Big(\frac{s}{B_1^{1/2}\sigma}\Big)\right\}\right]^{B_1/2} \leq \exp(-s^2/8).$$

Thus

$$\int_{S_1\leq|s|\leq S_2}\left|\frac{F^*_{B_1}(s)}{s}\right| ds \leq \frac{2}{S_1^2}\int_{S_1}^{S_2} s\exp(-s^2/8)\,ds \leq \frac{2^{13}}{B_1\sigma^3}. \tag{2.30}$$

Now,

$$\int_{S_1\leq|s|\leq S_2}\left|\frac{\Phi^*(s)}{s}\right| ds \leq \frac{2}{S_1^2}\int_0^{\infty} s\exp(-s^2/2)\,ds \leq \frac{2^{11}}{B_1\sigma^3}, \tag{2.31}$$

and

$$\int_{S_1\leq|s|\leq S_2}\left|\frac{P^*(s)}{s}\right| ds \leq \frac{1}{3S_1B_1^{1/2}\sigma}\int_0^{\infty} s^3\exp(-s^2/2)\,ds \leq \frac{2^6}{3\sqrt{2}B_1\sigma^3}. \tag{2.32}$$

To bound the final term, observe that, for all $|s| \in [S_1, S_2]$, since $(a+b)^2 \geq (a^2+b^2)/5$ for all $|a| \leq |b|/2$, we have

$$\begin{aligned}\int_{S_1\leq|s|\leq S_2}\left|\frac{Q^*(s)}{s}\right| ds &\leq \frac{1}{2\pi B_1^{1/2}\sigma}\int_{S_1\leq|s|\leq S_2} e^{-s^2/10}\int_{-\infty}^{\infty} e^{-2\pi^2 B_1\sigma^2 u^2/5}\,du\,ds \\ &\leq \frac{5}{2\sqrt{2}\pi B_1\sigma^3}.\end{aligned} \tag{2.33}$$

Finally, for $|s| \in [S_2, c_2 B_1^{1/2}\sigma]$, note that

$$\int_{S_2 \le |s| \le c_2 B_1^{1/2}\sigma} \left| \frac{\Phi^*(s) + P^*(s)}{s} \right| ds \le \frac{2}{S_2^2} \int_0^\infty s e^{-s^2/2}\, ds + \frac{1}{3 S_2 B_1^{1/2}\sigma} \int_0^\infty s^3 e^{-s^2/2}\, ds$$
$$\le \frac{1}{\pi^2 B_1 \sigma^3} \Big(1 + \frac{\pi}{3} \Big). \tag{2.34}$$

To bound the remaining terms, by substituting $s = B_1^{1/2}\sigma u$, we see that

$$\int_{S_2}^{c_2 B_1^{1/2}\sigma} \left| \frac{F_{B_1}^*(s) - Q_{B_1}^*(s)}{s} \right| ds = \int_\pi^{c_2} \left| \frac{F_{B_1}^*(B_1^{1/2}\sigma u) - Q_{B_1}^*(B_1^{1/2}\sigma u)}{u} \right| du$$
$$= \sum_{j=1}^{J} \int_{\pi(2j-1)}^{\pi(2j+1)} \left| \frac{F_{B_1}^*(B_1^{1/2}\sigma u) - Q_{B_1}^*(B_1^{1/2}\sigma u)}{u} \right| du$$
$$+ \int_{\pi(2J+1)}^{c_2} \left| \frac{F_{B_1}^*(B_1^{1/2}\sigma u) - Q_{B_1}^*(B_1^{1/2}\sigma u)}{u} \right| du, \tag{2.35}$$

where $J := \lfloor \frac{c_2 - \pi}{2\pi} \rfloor$. Let

$$I_j := \int_{\pi(2j-1)}^{\pi(2j+1)} \left| \frac{F_{B_1}^*(B_1^{1/2}\sigma u) - Q_{B_1}^*(B_1^{1/2}\sigma u)}{u} \right| du$$
$$= \int_{-\pi}^{\pi} \left| \frac{F_{B_1}^*\big(B_1^{1/2}\sigma(v + 2\pi j)\big) - Q_{B_1}^*\big(B_1^{1/2}\sigma(v + 2\pi j)\big)}{v + 2\pi j} \right| dv. \tag{2.36}$$

Observe that

$$F_{B_1}^*\big(B_1^{1/2}\sigma(v + 2\pi j)\big) = \Big[(1-\theta) \exp\big\{ -i(v + 2\pi j)\theta \big\} + \theta \exp\big\{ i(v + 2\pi j)(1-\theta) \big\} \Big]^{B_1}$$
$$= \exp(-i 2\pi B_1 j \theta) \big[(1-\theta) \exp(-iv\theta) + \theta \exp\{ iv(1-\theta) \} \big]^{B_1}$$
$$= \exp(-i 2\pi B_1 j \theta) F_{B_1}^*(B_1^{1/2}\sigma v).$$

Similarly,

$$Q_{B_1}^*\big(B_1^{1/2}\sigma(v + 2\pi j)\big)$$
$$= -\frac{(v + 2\pi j)}{2\pi} \sum_{l \in \mathbb{Z} \setminus \{0\}} \frac{\exp(i 2\pi B_1 l \theta)}{l} \exp\Big\{ -\frac{B_1 \sigma^2}{2} \big(v + 2\pi j + 2\pi l \big)^2 \Big\}$$
$$= \frac{(v + 2\pi j) \exp(-i 2\pi B_1 j \theta)}{2\pi j} \exp\Big(-\frac{B_1 \sigma^2 v^2}{2} \Big)$$
$$- \frac{(v + 2\pi j)}{2\pi} \sum_{l \in \mathbb{Z} \setminus \{0, -j\}} \frac{\exp(i 2\pi B_1 l \theta)}{l} \exp\Big\{ -\frac{B_1 \sigma^2}{2} \big(v + 2\pi j + 2\pi l \big)^2 \Big\}.$$

But, for $v \in [-\pi, \pi]$,

$$\left|\frac{1}{2\pi}\sum_{l\in\mathbb{Z}\setminus\{0,-j\}}\frac{e^{i2\pi B_1 l\theta}}{l}\exp\left\{-\frac{B_1\sigma^2}{2}\big(v+2\pi j+2\pi l\big)^2\right\}\right| \le \frac{1}{2\pi}\sum_{m\in\mathbb{Z}\setminus\{0\}} e^{-\frac{B_1\sigma^2}{2}(v+2\pi m)^2}$$
$$\le \frac{e^{-B_1\sigma^2 v^2/10}}{2\pi}\sum_{m\in\mathbb{Z}\setminus\{0\}} e^{-2\pi^2 B_1\sigma^2 m^2/5} \le \frac{e^{-B_1\sigma^2 v^2/10}}{\pi(e^{2\pi^2 B_1\sigma^2/5}-1)} \le \frac{5e^{-B_1\sigma^2 v^2/10}}{2\pi^3 B_1\sigma^2}.$$

It follows that

$$I_j \le \int_{-\pi}^{\pi}\left|\frac{F^*_{B_1}(B_1^{1/2}\sigma v) - \big(\frac{v}{2\pi j}+1\big)\exp\big(-\frac{B_1\sigma^2 v^2}{2}\big)}{v+2\pi j}\right| dv + \frac{5\sqrt{5}}{\sqrt{2}\pi^{5/2}B_1^{3/2}\sigma^3}. \tag{2.37}$$

Now

$$\int_{-\pi}^{\pi}\left|\frac{F^*_{B_1}(B_1^{1/2}\sigma v) - \exp\big(-\frac{B_1\sigma^2 v^2}{2}\big)}{v+2\pi j}\right| dv \le \frac{1}{\pi j B_1^{1/2}\sigma}\int_{-\pi B_1^{1/2}\sigma}^{\pi B_1^{1/2}\sigma}\left|F^*_{B_1}(u) - e^{-u^2/2}\right| du$$
$$= \frac{1}{\pi j B_1^{1/2}\sigma}\int_{-S_3}^{S_3}\left|F^*_{B_1}(u) - e^{-u^2/2}\right| du$$
$$+ \frac{1}{\pi j B_1^{1/2}\sigma}\int_{S_3\le|u|\le\pi B_1^{1/2}\sigma}\left|F^*_{B_1}(u) - e^{-u^2/2}\right| du, \tag{2.38}$$

where $S_3 := \frac{B_1^{1/2}\sigma}{5(2\theta^2-2\theta+1)} \ge S_1$. By Gnedenko and Kolmogorov (1954, Theorem 2, Section 40), we have that

$$\frac{1}{\pi j B_1^{1/2}\sigma}\int_{-S_3}^{S_3}\left|F^*_{B_1}(u) - e^{-u^2/2}\right| du \le \frac{7}{6\pi j B_1\sigma^2}\int_{-S_3}^{S_3}|u|^3 e^{-u^2/4}\, du \le \frac{56}{3\pi j B_1\sigma^2}. \tag{2.39}$$

Moreover,

$$\frac{1}{\pi j B_1^{1/2}\sigma}\int_{S_3\le|u|\le\pi B_1^{1/2}\sigma}\left|F^*_{B_1}(u) - e^{-u^2/2}\right| du \le \frac{2}{\pi j S_3 B_1^{1/2}\sigma}\int_0^{\infty} u(e^{-u^2/8} + e^{-u^2/2})\, du$$
$$\le \frac{50}{\pi j B_1\sigma^2}. \tag{2.40}$$

Finally,

$$\frac{1}{2\pi j}\int_{-\pi}^{\pi}\frac{|v|}{|v|+2\pi j}\exp\left(-\frac{B_1\sigma^2 v^2}{2}\right) dv \le \frac{1}{2\pi^2 j^2}\int_0^{\pi} v\exp\left(-\frac{B_1\sigma^2 v^2}{2}\right) dv$$
$$\le \frac{1}{2\pi^2 j^2 B_1\sigma^2}. \tag{2.41}$$

By (2.35), (2.36), (2.37), (2.38), (2.39), (2.40) and (2.41), it follows that

$$\int_{S_2\leq|s|\leq c_2 B_1^{1/2}\sigma}\left|\frac{F^*_{B_1}(s)-Q^*_{B_1}(s)}{s}\right|ds \leq \frac{10\sqrt{5}(J+1)}{\sqrt{2}\pi^{5/2}B_1^{3/2}\sigma^3}+\frac{140}{\pi B_1\sigma^2}\sum_{j=1}^{J+1}\frac{1}{j}$$
$$=\frac{10\sqrt{5}(J+1)}{\sqrt{2}\pi^{5/2}B_1^{3/2}\sigma^3}+\frac{140}{\pi B_1\sigma^2}\{1+\log(J+1)\}. \tag{2.42}$$

By (2.27), (2.28), (2.29), (2.30), (2.31), (2.32), (2.33), (2.34) and (2.42), we conclude that (2.26) holds. The result now follows from Theorem 2.6, by taking $c_1 = \frac{1}{B_1^{1/2}\sigma}$, $C_1 = \frac{1}{3B_1^{1/2}\sigma}$ and $S = c_2 B_1^{1/2}\sigma$ in that result. □

2.9 R code

Here we present the R code for the simulation experiments presented in Section 2.6.

```
setwd("~/RPEnsemble/Sims")
library(class,mvtnorm,MASS,Matrix,parallel,distr,ascrda,pamr,
   penalizedLDA,randomForest,e1071,kernlab)

##RandProjHaar: generates a random projection
RandProjHaar <- function(p=1000, d=10)
{
  R <- matrix(1/sqrt(p)*rnorm(p*d, 0, 1), p, d)
  R <- qr.Q(qr(R))[, 1:d]
  return(R)
}

##RPChoose: Chooses the best projection from a block of size B2
RPChoose <- function(B2, d, XTrain, YTrain, XTest, k = c(3,5,10,15,25)
   )
{
  n <- length(YTrain)
  p <- ncol(XTrain)
  w1 <- n
  w2 <- n
  w3 <- n
  for (j in 1:B2)
  {
    RP <- RandProjHaar(p, d)
    kcv.voteRP <- sapply(k, function(x){sum(knn.cv(XTrain%*%RP, YTrain
   , x) != YTrain, na.rm = TRUE)/n})
    weight.test1 <- min(kcv.voteRP)
    if (weight.test1 <= w1)
```

```
    {
      w1 <- weight.test1
      RP1 <- RP
      k1 <- order(kcv.voteRP)[1]
    }
    weight.test2 <-  mean(predict(lda(x = XTrain%*%RP, grouping =
   YTrain), XTrain%*%RP)$class != YTrain, na.rm = TRUE)
    if (weight.test2 <= w2)
    {
      w2 <- weight.test2
      RP2 <- RP
    }
    weight.test3 <- mean(qda(x = XTrain%*%RP, grouping = YTrain, CV =
   TRUE)$class != YTrain, na.rm = TRUE)
    if (weight.test3 <= w3)
    {
      w3 <- weight.test3
      RP3 <- RP
    }
  }
Cutoff.val <- c(as.numeric(knn.cv(XTrain%*%RP1, YTrain, k1)), as.
   numeric(predict(lda(x = XTrain%*%RP2,grouping = YTrain), XTrain%*%
   RP2)$class), as.numeric(qda(x = XTrain%*%RP3,grouping = YTrain, CV
    = TRUE)$class))
class.vote <- c(as.numeric(knn(XTrain%*%RP1, XTest%*%RP1, YTrain, k1))
   , as.numeric(predict(lda(x = XTrain%*%RP2, grouping = YTrain),
   XTest%*%RP2)$class), as.numeric(predict(qda(x = XTrain%*%RP3,
   grouping = YTrain), XTest%*%RP3)$class))
return(c(Cutoff.val, class.vote))
}

## RPclassifer: averages over B1 carefully chosen random projections
   in parallel to classify the test set
RPClassifier <- function(XTrain, YTrain, XTest, YTest, d, B1 = 100, B2
    = 100, k = c(3,5,9,15,25))
{
  n <- length(YTrain)
  n.test <- length(YTest)
  p = ncol(XTrain)
  n1 <- table(YTrain)[[1]]
  n2 <- table(YTrain)[[2]]
  p1 <- n1/(n1+n2)
  p2 <- 1-p1
  RP.out <- simplify2array(mclapply(rep(1,B1), function(x){return(
   RPChoose(B2, d, XTrain, YTrain, XTest, k))}, mc.cores = 4))
  errRP <- rep(0,3)
```

```
  for (b in 1:3)
  {
    b1 <- (b-1)*n
    b2 <- 3*n + (b-1)*n.test
    Cutoff.val <- RP.out[b1+1:n, ]
    Class.vote <- RP.out[b2+1:n.test, ]
    vote1 <- rowMeans(Cutoff.val[1:n1, ], na.rm = TRUE)
    vote2 <- rowMeans(Cutoff.val[n1+1:n2, ], na.rm = TRUE)
    errecdfm <- function(x){
      p1*ecdf(vote1)(x) + (1-p1)*(1-ecdf(vote2)(x))
    }
    errecdfM <- function(x){
      p1*ecdf(vote1)(-x) + (1-p1)*(1-ecdf(vote2)(-x))
    }
    alpham <- optimise(errecdfm,c(1,2),maximum=TRUE)$maximum
    alphaM <- optimise(errecdfM,c(-2,-1),maximum=TRUE)$maximum
    alpha <- (alpham-alphaM)/2
    vote <- rowMeans(Class.vote, na.rm = TRUE)
    Class  <- 1 + as.numeric(vote > alpha)
    errRP[b]  <- 100*mean(Class != YTest, na.rm = TRUE)
  }
return(errRP)
}

##MainSim: RP Ensemble simulations
MainSim <- function(Model.No, n.train = 50, n.test = 1000, n.reps =
   100, d = 5, B1 = 100, B2 = 100,  k = c(3,5,9,15,25))
{
  Risk <- NULL
  for (i in 1:n.reps)
  {
    set.seed(100 + i)
    data.train <- Model(Model.No, n.train, p = 50, s0 = 1, Pi = 1/2)
    data.test <-  Model(Model.No, n.test, p = 50, s0 = 1, Pi = 1/2)
    if(i==1) Risk <-  RPClassifier(data.train$x, data.train$y, data.
   test$x, data.test$y, d = d, B1 = B1, B2 = B2)
    else Risk <- rbind(Risk, RPClassifier(data.train$x, data.train$y,
   data.test$x, data.test$y, d = d, B1 = B1, B2 = B2))
  }
return(Risk)
}

##CompSim: comparitor simulations
CompSim <- function(Model.No, n.train = 50, p = 50, s0 = 1, prior = 1/
   2, n.test = 1000, n.reps = 100, k = c(3,5,9,15,25))
{
```

```
for (i in 1:n.reps)
{
  set.seed(100 + i)
  data.train <- Model(Model.No, n.train, p, s0, prior)
  data.test <- Model(Model.No, n.test, p, s0, prior)
  n.min <- min(table(data.train$y))

  errLDA <- NA
  if (n.train > p) errLDA <- 100*mean(predict(lda(x = data.train$x,
 grouping = data.train$y), data.test$x)$class != data.test$y, na.rm
  = TRUE)

  errQDA <- NA
  if (n.min > p) errQDA <- 100*mean(predict(qda(x = data.train$x,
 grouping = data.train$y), data.test$x)$class != data.test$y, na.rm
  = TRUE)

  errknn <- NA
  kcv <- sapply(k,function(x){sum(knn.cv(data.train$x, data.train$y,
  x) != data.train$x)})
  errknn <- 100*mean(knn(data.train$x, data.test$x, data.train$y, k
= k[which.min(kcv)]) != data.test$y, na.rm = TRUE)

  err1PEN <- NA
  cv.out1 <- PenalizedLDA.cv(data.train$x, data.train$y, lambdas=c
(0.055,0.06,0.065,.07,.075,0.08,.085,0.09,0.095))
  out1  <- PenalizedLDA(data.train$x, data.train$y, data.test$x,
lambda = cv.out1$bestlambda, K = cv.out1$bestK)
  err1PEN <- 100*mean(out1$ypred[,1] - data.test$y != 0, na.rm =
TRUE)

  err1NSC <- NA
  Trainout1 <- pamr.train(list(x = t(data.train$x), y = data.train$y
))
  CV.out1NSC <- pamr.cv(Trainout1, list(x = t(data.train$x), y =
data.train$y), nfold = 10)
  out1NSC <- pamr.predict(Trainout1, t(data.test$x), threshold = CV.
out1NSC$threshold[which.min(CV.out1NSC$error)], type = c("class"))
  err1NSC <- 100*mean(out1NSC != data.test$y, na.rm = TRUE)

  err1SCRDA <- NA
  out1SCRDA <- ascrda(data.train$x, data.train$y, data.test$x, data.
test$y, SCRDAmethod = "SCRDA")
  err1SCRDA <- as.numeric(out1SCRDA$SCRDA)*100

  err1IR <- NA
```

```
    err1IR <- FitDLDA(data.train$x, data.train$y, data.test$x, data.
   test$y)$Err*100

    errRandForest <- NA
    RandForest <- randomForest(x = data.train$x, y = factor(data.train
   $y), xtest = data.test$x, ytest = factor(data.test$y), ntree =
   1000, mtry = sqrt(p), replace = TRUE, classwt = c(n1/n.train, n2/n
   .train), cutoff = c(0.5,0.5), sampsize = n.train, nodesize = 1,
   keep.forest= TRUE)
    errRandForest <- 100*mean(as.numeric(predict(RandForest, newdata =
    data.test$x)) != data.test$y, na.rm = TRUE)

    errSVMRad <- NA
    SVMRad <- svm(x = data.train$x, y = factor(data.train$y), kernel =
    "radial", gamma = 1/p, cost = 1, class.weights = list("1" = n1/n.
   train, "2" =  n2/n.train), cachesize = 40, tolerance = 0.001,
   epsilon = 0.1, shrinking = TRUE, cross = 0, probability = FALSE,
   fitted = TRUE, na.action = na.omit)
    errSVMRad <- 100*mean(as.numeric(predict(SVMRad, newdata = data.
   test$x)) != data.test$y, na.rm = TRUE)

    errSVMLin <- NA
    SVMLin <- svm(x = data.train$x, y = factor(data.train$y), kernel =
    "linear", gamma = 1/p, cost = 1, class.weights = list("1" = n1/n.
   train, "2" =  n2/n.train), cachesize = 40, tolerance = 0.001,
   epsilon = 0.1, shrinking = TRUE, cross = 0, probability = FALSE,
   fitted = TRUE, na.action = na.omit)
    errSVMLin <- 100*mean(as.numeric(predict(SVMLin, newdata = data.
   test$x)) != data.test$y, na.rm = TRUE)

    errGPRad <- NA
    GPRad <- gausspr(x = data.train$x, y = factor(data.train$y),
   scaled = FALSE, type= "classification", kernel="rbfdot", kpar="
   automatic", var=1, variance.model = FALSE, tol=0.0005, cross=0,
   fit=FALSE, na.action = na.omit)
    errGPRad <- 100*mean(1 + (predict(GPRad, newdata = data.test$x,
   type = "probabilities", coupler = "pkpd")[,2]>0.5) != data.test$y)

    risknew <-  c(errLDA, errQDA, errknn,errRandForest, errSVMRad,
   errSVMLin, errGPRad, errGPLin, err1PEN, err1NSC, err1SCRDA, err1IR
   )
    if(i==1) Risk <- risknew
    else Risk <- rbind(Risk, risknew)
  }
return(Risk)
}
```

```
##Model: Generates data
Model <- function(Model.No, n, p, s0, Pi = 1/2)
{
  if (Model.No == 1)
  {
    Y1 <- rmultinom(1, n, c(Pi, 1-Pi))
    Y <- c(rep(1, Y1[1,1]), rep(2, Y1[2,1]))
    mu <- rep(1/8, p)
    D <- DExp(1)
    X1 <- cbind(matrix(r(D)(Y1[1,1]*p), Y1[1,1], p))
    X2 <- mvrnorm(Y1[2,1], mu, diag(p))
    X  <- rbind(X1, X2)
  }
  if (Model.No == 2)
  {
    Y1 <- rmultinom(1, n, c(Pi, 1-Pi))
    Y <- c(rep(1, Y1[1,1]),rep(2, Y1[2,1]))
    mu <- c(rep(2, 5), rep(0, p-5))
    U1 <- rchisq(Y1[1, 1], 1)
    U2 <- rchisq(Y1[2, 1], 2)
    Sigma1 <- diag(p)
    Sigma2 <- 0.5*diag(p) + 0.5*c(rep(1, 5), rep(0, p-5))%*%t(c(rep(1,
    5) ,rep(0, p-5))) + 0.5*diag(c(rep(0, 5), rep(1, p-5)))
    X1 <- mvrnorm(Y1[1, 1], rep(0, p), Sigma1)/sqrt(U1/1)
    X2 <- t(mu + t(mvrnorm(Y1[2, 1], rep(0, p), Sigma2)/sqrt(U2/2)))
    X  <- rbind(X1, X2)
  }
  if (Model.No == 3)
  {
    Y1 <- rmultinom(1, n, c(Pi, 1-Pi))
    Y <- c(rep(1, Y1[1, 1]), rep(2, Y1[2, 1]))
    Y11 <- rmultinom(1, Y1[1, 1], c(1/2, 1/2))
    mu <- c(rep(1, 5), rep(0, p-5))
    Sigma <- diag(p)
    X1 <- rbind(t(matrix(mu/2, p, Y11[1, 1])), t(matrix(mu/2, p, Y11
   [2, 1]))) + mvrnorm(Y1[1, 1], rep(0, p), Sigma)
    X2 <- cbind(matrix(rcauchy(Y1[2, 1]*5), Y1[2, 1], 5), matrix(rnorm
   (Y1[2, 1]*(p-5), 0, 1), Y1[2, 1], p-5))
    X <- rbind(X1, X2)
  }
  if (Model.No == 4)
  {
    load("R.RData")
    Y1 <- rmultinom(1, n, c(Pi, 1-Pi))
    Y <- c(rep(1, Y1[1, 1]), rep(2, Y1[2, 1]))
```

```
    mu <- c(rep(1, 3), rep(0, p-3))
    Sigma1 <- 0.5*diag(c(rep(1, 3), rep(0, p-3))) + 0.5*c(rep(1, 3),
   rep(0, p-3))%*%t(c(rep(1, 3), rep(0, p-3))) + 0.5*diag(c(rep(0, 3)
   , rep(1, p-3))) + 0.5*c(rep(0, 3), rep(1, p-3))%*%t(c(rep(0, 3),
   rep(1, p-3)))
    Sigma2 <- 1.5*diag(c(rep(1, 3), rep(0, p-3))) + 0.5*c(rep(1, 3),
   rep(0, p-3))%*%t(c(rep(1, 3), rep(0, p-3))) + 0.5*diag(c(rep(0, 3)
   , rep(1, p-3))) + 0.5*c(rep(0, 3), rep(1, p-3))%*%t(c(rep(0, 3),
   rep(1, p-3)))
    X1 <- mvrnorm(Y1[1, 1], R%*%rep(0, p), R%*%Sigma1%*%t(R))
    X2 <- mvrnorm(Y1[2, 1], R%*%mu, R%*%Sigma2%*%t(R))
    X <- rbind(X1, X2)
  }
  if (Model.No == 5)
  {
    mu <- c(rep(0, p-3))
    EX <- mvrnorm(n, rep(0, p-3), diag(p-3))
    Z2 <- matrix(runif(3*n, -1, 1), n, 3)
    Y1 <- sum(diag(Z2%*%t(Z2)) >= 1)
    Y <- c(rep(1, Y1), rep(2, n-Y1))
    X1 <- Z2[diag(Z2%*%t(Z2)) >= 1, ]
    X2 <- Z2[diag(Z2%*%t(Z2)) < 1, ]
    X <- cbind(rbind(X1, X2), EX)
  }
return(list(x=X, y=Y))
}

##Settings <- c(Model.No, n, p, prior, ntest, nreps, d, B1, B2)
##Settings <- read.csv("Settings.txt", header = T)
Settings <- rbind(c(1, 50, 50, 0.5, 1000, 100, 2, 100, 100), c(1, 50,
   50, 0.5, 1000, 100, 5, 100, 100))
for (Job in 1:2)
{
  out <- MainSim(Model.No = Settings[Job, 1], n.train = Settings[Job,
   2], p = Settings[Job, 3], prior = Settings[Job, 4], n.test =
   Settings[Job, 5], n.reps = Settings[Job, 6], d = Settings[Job, 7],
    B1 = Settings[Job, 8], B2 = Settings[Job, 9])
  save(out, file = paste("Risk-",Job,".RData", sep=""))
}
for (Job in 1)
{
  outcomp <- CompSim(Model.No = Settings[Job, 1], n.train = Settings[
   Job, 2], p = Settings[Job, 3], prior = Settings[Job, 4], n.test =
   Settings[Job, 5], n.reps = Settings[Job, 6])
  save(outcomp, file = paste("RiskComp-",Job,".RData", sep=""))
}
```

Chapter 3

Semi-supervised tail adaptive nearest neighbour classification

3.1 Introduction

In this chapter, we propose a semi-supervised k-nearest neighbour classifier, where the number of neighbours considered varies with the location of the test point. More precisely, we first estimate the marginal density of the features using a large unlabelled training data set, then let k depend on this estimate at the test point, using fewer neighbours when the density is small. The method is motivated by a new asymptotic expansion for the global excess risk of the standard k-nearest neighbour classifier. This expansion elucidates conditions under which the dominant contribution to the excess risk comes from the locus of points at which each class label is equally likely to occur, as well as situations where the dominant contribution comes from the tails of the marginal distribution of the features. We show further that the proposed semi-supervised classifier exploits a local bias-variance trade-off. As a result, the tail excess risk is shown to be negligible when the features have more than four finite moments, regardless of their dimension d (for the standard k-nearest neighbour classifier, our theory requires $d \geq 5$ and more than $4d/(d-4)$ finite moments).

Introduced by Fix and Hodges (1951), the k-nearest neighbour ($k$nn) classifier assigns the test point according to a majority vote over the classes of its k nearest points in the training set. While this simple and intuitive method has become extremely popular in practical problems, it was not until recently that detailed understanding of its error properties was known (Hall, Park and Samworth, 2008; Samworth, 2012a). Even then the expansion of the excess risk is restricted to a compact set, ignoring the tail of the distribution.

The first goal in this chapter is to characterise the error properties of the standard k-nearest neighbour classifier in the tail of the feature vector distribution, when these

vectors take values in $\mathbb{R}^d$. Intuitively, points in the tail will be harder to classify since there are typically fewer training data points in that region. On the other hand such points will rarely be observed, and may contribute little to the overall risk even if classified incorrectly. In Theorem 3.2 in Section 3.3 we investigate and characterise this trade-off. It is shown that, under some regularity conditions, for $d \geq 5$, if the features have more than $4d/(d-4)$ finite moments, then the excess risk in the tail of the k-nearest neighbour classifier is asymptotically smaller than the risk in the body of the distribution. In this case, we derive an asymptotic expansion for the *global* excess risk. However, when $d \leq 4$, or when moment condition above is not satisfied, the error in the tail may dominate.

The results discussed above motivate a modification of the $k$nn classifier in semi-supervised classification settings. We propose to allow the choice of k to depend on (an estimate of) the feature vector density $\bar{f}$ at the test point. We argue that, by using fewer neighbours in low density regions, we are able to achieve a better balance in the local bias-variance trade-off. In particular, using an oracle, local choice of k that depends on $\bar{f}$, and under regularity conditions, we show that the excess risk over $\mathbb{R}^d$ is $O(n^{-4/(d+4)})$ provided that the feature vectors have more than four finite moments. By contrast, our theory for the standard $k$nn classifier with a global choice of k requires that $d \geq 5$ and the feature vectors have more than $4d/(d-4)$ finite moments. Of course, in practice $\bar{f}$ is unknown, but in a semi-supervised setting with m additional, independent, unlabelled observations, we can estimate it by $\hat{f}_m$, say. Provided $m/n^{2+d/2} \to \infty$, we show that the tail-adaptive semi-supervised classifier mimics the asymptotic performance of the oracle.

The local classifier is similar in spirit to a neighbourhood classifier (see, for example, Owen, 1984), where classification is made according to a majority vote over the classes of the points in a ball of fixed radius about the test point. On the other hand, there are few results on the topic of classification in the tails. Indeed, asymptotic expansions for the excess risk of *plug-in* type classifiers usually require that the feature vectors are compactly supported (Mammen and Tsybakov, 1999; Audibert and Tsybakov, 2007; Biau, Cérou and Guyader, 2010), or at least the expansion is restricted to a compact set (Hall and Samworth, 2005). Hall, Park and Samworth (2008) showed that, under weak regularity conditions, the excess risk of the k-nearest neighbour classifier over a compact set $\mathcal{R}$ is given by

$$B_{1,\mathcal{R}}\frac{1}{k} + B_{2,\mathcal{R}}\Big(\frac{k}{n}\Big)^{4/d} + o\big\{1/k + (k/n)^{4/d}\big\} \tag{3.1}$$

as $n \to \infty$; see also Samworth (2012a). Let $(X, Y) \in \mathbb{R}^d \times \{1, 2\}$ denote a generic feature vector–class label pair, and let $\eta(x) := \mathbb{P}(Y = 1|X = x)$ denote the regression function. The constants $B_{1,\mathcal{R}}$ and $B_{2,\mathcal{R}}$, given explicitly in Samworth (2012a), depend

only on the behaviour of the joint distribution of (X, Y) in a neighbourhood of the Bayes decision boundary, where $\eta(x) = 1/2$. The expansion in (3.1) can be interpreted as a bias-variance decomposition of an estimate of η. A simple calculation shows that the optimal k, which balances the two dominant terms to minimise the risk, is given by

$$k^*_{\mathcal{R}} = \Big\lfloor \Big(\frac{dB_{1,\mathcal{R}}}{4B_{2,\mathcal{R}}} \Big)^{d/(d+4)} n^{4/(d+4)} \Big\rfloor. \tag{3.2}$$

The excess risk of the $k^*_{\mathcal{R}}$-nearest neighbour classifier over $\mathcal{R}$ is $O(n^{-4/(d+4)})$. Samworth (2012a) shows further that by using a weighted classifier – that is by assigning increasingly distant neighbours a smaller weight – one can guarantee improvements in the constant preceding the leading order term in the asymptotic expansion. The improvement is at least 5% when $d \leq 15$, and most significant when $d = 4$. As d increases, however, the benefit becomes negligible.

Hall and Kang (2005) study the tail error properties of a kernel classifier for univariate data. Their method first computes kernel estimates for the class conditional densities, then classifies according to which density estimate, weighted by the prior probability of each class, is larger. However, both estimates may be zero, for instance for points beyond the maximum order statistic of the training sample. In this case, the test point is classified according to which density is first non-zero as one looks back toward the body of the distribution. It is shown that, under certain regularity conditions, the contribution to the excess risk from the tail is of smaller order than the contribution from the body of the distribution. However, these regularity conditions exclude, for example, Pareto-type tails. For instance, following an example from Hall and Kang (2005), suppose that, for large x, one class has density $ax^{-\alpha}$, while the other has density $bx^{-\beta}$, where $a, b > 0$ and $1 < \alpha < \beta < \alpha + 1 < \infty$. Then the excess risk from the right tail is of larger order than that in the body of the distribution. We will see later that the $k$nn classifier does not suffer the same disadvantage and that, in the example above – in fact, the contribution to the excess risk of the $k$nn classifier is $O(n^{-M})$ for every $M > 0$.

Gedat, Klein and Marteau (2014) derive global rates of convergence for the k-nearest neighbour classifier, when η is Lipschitz and the well-known *margin assumption* of Mammen and Tsybakov (1999); Tsybakov (2004) is satisfied. Under the further condition that $\mathbb{P}\{\bar{f}(X) < \delta\} < \delta$, as $\delta \to 0$, it is shown that the global risk of $k$nn classifier, with $k = \lfloor n^{\frac{2}{3+\alpha+d}} \rfloor$, is $O(n^{-\frac{1+\alpha}{3+\alpha+d}})$. This is a slower rate than the $O(n^{-\frac{1+\alpha}{2+d}})$ rate that one can expect under the same conditions when considering the risk over a compact set.

The remainder of this chapter is organised as follows. After introducing our notation and giving a preliminary result in Section 3.2, we present in Section 3.3 our main results for the standard $k$nn classifier. This leads on, in Section 3.4, to our study of the

semi-supervised setting, where we derive asymptotic results of the excess risk of our tail-adaptive classifier. We illustrate the finite-sample benefits of the adaptive classifiers over the standard $k$nn classifier in a simulation study in Section 3.5. Technical arguments are deferred to the appendix.

Finally we fix here some notation used throughout this chapter. Let $\|\cdot\|$ denote the Euclidean norm and, for $\delta > 0$ and $x \in \mathbb{R}^d$, let $B_\delta(x) := \{z \in \mathbb{R}^d : \|x - z\| \leq \delta\}$ denote the closed ball of radius δ centred at x. Let $a_d := \frac{2\pi^{d/2}}{d\Gamma(d/2)}$ denote the d-dimensional Lebesgue measure of $B_1(0)$. For a real-valued function g defined on $A \subseteq \mathbb{R}^d$ that is twice differentiable at x, write $\dot{g}(x) = (g_1(x), \ldots, g_d(x))^T$ and $\ddot{g}(x) = \big(g_{jk}(x)\big)$ for its gradient vector and Hessian matrix at x, and let $\|g\|_\infty = \sup_{x \in A} |g(x)|$. Let $\|\cdot\|_{\mathrm{op}}$ denote the operator norm of a matrix.

3.2 Statistical setting

We recall the semi-supervised classification setting outlined in the main introduction. Let $(X, Y), (X_1, Y_1), \ldots, (X_{n+m}, Y_{n+m})$ be independent random pairs taking values in $\mathbb{R}^d \times \{1, 2\}$, each with joint distribution P. Let $\pi_r := \mathbb{P}(Y = r)$, for $r = 1, 2$, and $X|Y = r \sim P_r$, for $r = 1, 2$, where P_r is a probability measure on $\mathbb{R}^d$. Define the regression function $\eta(x) := \mathbb{P}(Y = 1|X = x)$ and let $P_X := \pi_1 P_1 + \pi_2 P_2$ denote the marginal distribution of X.

We observe *labelled training data*, $\mathcal{T}_n := \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, and *unlabelled training data*, $\mathcal{T}'_m := \{X_{n+1}, \ldots, X_{n+m}\}$, and are presented with the task of assigning the *test point* X to either class 1 or 2.

A *classifier* is a Borel measurable function $C : \mathbb{R}^d \to \{1, 2\}$, with the interpretation that C assigns $x \in \mathbb{R}^d$ to the class $C(x)$. Given a Borel measurable set $\mathcal{R} \subseteq \mathbb{R}^d$, the misclassification rate, or *risk*, over $\mathcal{R}$ is

$$R_{\mathcal{R}}(C) := \mathbb{P}[\{C(X) \neq Y\} \cap \{X \in \mathcal{R}\}]. \tag{3.3}$$

Here, the set $\mathcal{R}$ may be the whole of $\mathbb{R}^d$, in which case the subscript $\mathcal{R}$ will be dropped, or a particular subset of $\mathbb{R}^d$ of interest. The *Bayes classifier*

$$C^{\mathrm{Bayes}}(x) := \begin{cases} 1 & \text{if } \eta(x) \geq 1/2; \\ 2 & \text{otherwise,} \end{cases} \tag{3.4}$$

minimises the risk over any region $\mathcal{R}$ (Devroye et al., 1996, p. 20). Thus, the performance of a classifier C is measured via its (non-negative) *excess risk*, $R_{\mathcal{R}}(C) - R_{\mathcal{R}}(C^{\mathrm{Bayes}})$. A classifier $\hat{C}_n$, based on the training data $\mathcal{T}_n$, is said to be *consistent* if the excess risk converges to zero as $n \to \infty$, but we will also be interested in a more

precise description of the asymptotic behaviour of the excess risk.

We can now formally define the local-k-nearest neighbour classifier, which allows the number of neighbours considered to vary depending on the location of the test point. Suppose $k_{\mathrm{L}} : \mathbb{R}^d \to \{1, \ldots, n\}$ is measurable. Given the test point $x \in \mathbb{R}^d$, let $(X_{(1)}, Y_{(1)}), \ldots, (X_{(n)}, Y_{(n)})$ be a reordering of the training data such that $\|X_{(1)} - x\| \leq \cdots \leq \|X_{(n)} - x\|$. (When a distinction needs to be made, we also write $X_{(i)}(x)$ for the ith nearest neighbour of x.) Let $\hat{S}_n(x) := k_{\mathrm{L}}(x)^{-1} \sum_{i=1}^{k_{\mathrm{L}}(x)} \mathbb{1}_{\{Y_{(i)}=1\}}$. Then the *local-$k$-nearest neighbour classifier* (k_{L}nn) is defined to be

$$\hat{C}_n^{k_{\mathrm{L}}\mathrm{nn}}(x) := \begin{cases} 1 & \text{if } \hat{S}_n(x) \geq 1/2; \\ 2 & \text{otherwise.} \end{cases} \tag{3.5}$$

Given $k \in \{1, \ldots, n\}$, let k_0 denote the constant function $k_0(x) := k$ for all $x \in \mathbb{R}^d$. Using $k_{\mathrm{L}} = k_0$ the definition above reduces to the standard *k-nearest neighbour classifier* ($k$nn), and we will write $\hat{C}_n^{k\mathrm{nn}}$ in place of $\hat{C}_n^{k_0\mathrm{nn}}$.

For $\beta \in (0, 1)$, let

$$K_{\beta,0} := \{\lceil \log^4 n \rceil, \lceil \log^4 n \rceil + 1, \ldots, \lfloor n^{1-\beta} \rfloor\} \tag{3.6}$$

denote a region of values of k that will be of interest to us. Note that $K_{\beta_1,0} \supset K_{\beta_2,0}$, for $\beta_1 < \beta_2$. Moreover, when β is small, the upper and lower bounds are only a very slightly stronger requirement than Stone's conditions for consistency that $k = k_n \to \infty$, $k_n/n \to 0$ as $n \to \infty$.

3.2.1 Preliminary result

In Proposition 3.1 below we show that, if η is bounded away from $1/2$ in the tails, then the $k$nn classifier performs very well in that region.

Proposition 3.1. *Let $\delta > 0$ and suppose that $\mathcal{R} \subseteq \mathbb{R}^d$ is a bounded, open, convex set with the property that $\inf_{x \in \partial\mathcal{R}} P_X\big(B_\delta(x)\big) > 0$. Now suppose that $\mathcal{R}_0 \subseteq \mathcal{R}$ is such that $x \in \mathcal{R}$ whenever $\|x - w\| \leq 2\delta$ for some $w \in \mathcal{R}_0$, and such that*

$$\inf_{x \in \mathcal{R}_0^c} \eta(x) - \frac{1}{2} > \epsilon \tag{3.7}$$

for some $\epsilon \in (0, 1/2)$. Then for each $\beta \in (0, 1)$ and each $M > 0$,

$$\sup_{k \in K_{\beta,0}} R_{\mathcal{R}^c}(\hat{C}_n^{k\mathrm{nn}}) - R_{\mathcal{R}^c}(C^{\mathrm{Bayes}}) = O(n^{-M}) \tag{3.8}$$

as $n \to \infty$.

Remark The conclusion in (3.8) also holds if (3.7) is replaced with the condition that there exists $\epsilon \in (0, 1/2)$ such that $1/2 - \sup_{x \in \mathcal{R}_0^c} \eta(x) > \epsilon$.

When X is univariate, the assumption that η is bounded away from $1/2$ in the tail is relatively weak. Indeed, for the particular case described in the introduction, where P_1 and P_2 have Lebesgue densities f_1 and f_2 satisfying $f_1(x) \sim ax^{-\alpha}$ and $f_2(x) \sim bx^{-\beta}$ as $x \to \infty$, with $1 < \alpha < \beta < \alpha + 1 < \infty$, we in fact have that $\eta(x) \to 1$ as $x \to \infty$. It follows that, unlike the kernel-based classifiers studied by Hall and Kang (2005), the $k$nn classifier performs well in the tails in this case.

For multivariate X, the assumption in (3.7) is rather strong; in particular, it implies that $\{x \in \mathbb{R}^d : \eta(x) = 1/2\}$ is bounded. In Section 3.3 below, therefore, we consider the global risk of the k-nearest neighbour classifier in settings where η is not bounded away from $1/2$ in the tails.

3.3 Global risk of the k-nearest neighbour classifier

Our analysis will make use of the following assumptions:

(B.1) The probability measures P_1 and P_2 are absolutely continuous with respect to Lebesgue measure with twice continuously differentiable Radon–Nikodym derivatives f_1 and f_2, respectively, satisfying $\bar{f}(x) := \pi_1 f_1(x) + \pi_2 f_2(x) > 0$ for all $x \in \mathbb{R}^d$.

(B.2) The set $\mathcal{S} := \{x \in \mathbb{R}^d : \eta(x) = 1/2\}$ is non-empty with $\inf_{x_0 \in \mathcal{S}} \|\dot{\eta}(x_0)\| > 0$ and $\sup_{x_0 \in \mathcal{S}} a(x_0)^2 < \infty$, where the function a is given in (3.9) below. Furthermore, $\int_{\mathcal{S}} \bar{f}(x_0)\, d\mathrm{Vol}^{d-1}(x_0) < \infty$, where Vol^{d-1} denotes the natural $(d-1)$-dimensional volume measure that $\mathcal{S}$ inherits as a subset of $\mathbb{R}^d$.

(B.3) There exists $\delta_0 > 0$, such that

$$\sup_{\delta \in (0,\delta_0)} \sup_{x : \bar{f}(x) \geq \delta \log^{2d}(1/\delta)} \sup_{\|u\| \leq \log^{-2}(1/\delta)} \frac{\|\ddot{\bar{f}}(x+u)\|_{\mathrm{op}}}{\bar{f}(x) \log^2(2\|\bar{f}\|_\infty / \bar{f}(x))} < \infty.$$

(B.4)(ρ) We have that $\int_{\mathbb{R}^d} \|x\|^\rho dP_X(x) < \infty$ and $\int_{\mathcal{S}} \bar{f}(x_0)^{d/(\rho+d)}\, d\mathrm{Vol}^{d-1}(x_0) < \infty$.

The density assumption in **(B.1)** serves several purposes. First, it enables us to define the tail of the distribution as the region where $\bar{f}$ is smaller than some threshold. The assumption that $\bar{f}$ is supported on $\mathbb{R}^d$ finesses the problem of classification near the boundary of the support of $\bar{f}$. The behaviour of f_1 and f_2 in a neighbourhood of $\mathcal{S}$ is crucial for deriving an asymptotic expansion for the excess risk. The continuity of f_1 and f_2 ensures that η is continuous, so in particular it cannot jump discontinuously

across 1/2. Moreover, the smoothness of η and the lower bound on its derivative on $\mathcal{S}$, implies that $\mathcal{S}$ is a $(d-1)$-dimensional manifold, and facilitates the definition of the Vol^{d-1} measure. The assumption on the function

$$a(x) := \frac{\sum_{j=1}^{d}\{\eta_j(x)\bar{f}_j(x) + \frac{1}{2}\eta_{jj}(x)\bar{f}(x)\}}{(d+2)a_d^{2/d}\bar{f}(x)} \tag{3.9}$$

allows us to approximate the bias and variance of $\hat{S}_n(x)$, when x is in a neighbourhood of $\mathcal{S}$, by the first and second order terms in their asymptotic expansions. The last part of assumption **(B.2)** requires the restriction of the density to the set $\mathcal{S}$ to decay appropriately in the tails.

Assumption **(B.3)** is a technical condition on the second derivatives of $\bar{f}$, which means that the P_X-measure of a small ball at x is well approximated by taking the density to be constant over the ball. It is satisfied, for instance, by densities of the form, $\bar{f}(x) \propto \exp(-\|x\|^b)$, for $b > 0$, and $\bar{f}(x) \propto (1 + \|x\|)^{-(d+c)}$, for $c > 0$. The moment assumption in **(B.4)**(ρ) is used to bound the P_X-measure of the region where $\bar{f}$ is small.

We are now in a position to present our asymptotic expansion for the global excess risk of the standard $k$nn classifier.

Theorem 3.2. *Assume* ***(B.1)***, ***(B.2)***, ***(B.3)*** *and* ***(B.4)***(ρ).

(i) Suppose that $d \geq 5$ and $\rho > \frac{4d}{d-4}$. Then for each $\beta \in (0,1)$,

$$R(\hat{C}_n^{k\mathrm{nn}}) - R(C^{\mathrm{Bayes}}) = \Big\{\frac{B_1}{k} + B_2\Big(\frac{k}{n}\Big)^{4/d}\Big\}\{1 + o(1)\} \tag{3.10}$$

as $n \to \infty$, uniformly for $k \in K_{\beta,0}$, where B_1 and B_2 are given in (3.12).

(ii) Suppose that either $d \leq 4$, or, $d \geq 5$ and $\rho \leq \frac{4d}{d-4}$. Then for each $\beta \in (0,1)$,

$$R(\hat{C}_n^{k\mathrm{nn}}) - R(C^{\mathrm{Bayes}}) = \frac{B_1}{k} + o\Big(\frac{1}{k} + \Big(\frac{k}{n}\Big)^{\frac{\rho}{\rho+d}}\log^{2d\rho/(\rho+d)} n\Big) \tag{3.11}$$

as $n \to \infty$, uniformly for $k \in K_{\beta,0}$.

We see in Theorem 3.2, part (i) that, for $d \geq 5$, if P_X has enough moments then the main contribution to the risk arises from neighbourhood of the Bayes decision boundary in the body of the distribution, where $\bar{f}(x) \geq \frac{k}{n}\log^{2d}(n/k)$. Furthermore, due to the moment condition, the excess risk over $\big\{x \in \mathbb{R}^d : \bar{f}(x) < \frac{k}{n}\log^{2d}(n/k)\big\}$ is negligible in comparison. In this case we are able to give the dominant term in the asymptotic expansion of the global excess risk, where the integrability condition ensures that the

constants

$$B_1 := \int_{\mathcal{S}} \frac{\bar{f}(x_0)}{4\|\dot{\eta}(x_0)\|}\, d\mathrm{Vol}^{d-1}(x_0) \text{ and } B_2 := \int_{\mathcal{S}} \frac{\bar{f}(x_0)^{1-4/d}}{\|\dot{\eta}(x_0)\|} a(x_0)^2\, d\mathrm{Vol}^{d-1}(x_0), \quad (3.12)$$

are finite.

On the other hand, in part (ii), it is possible that the main contribution to the excess risk is coming from the region where $\bar{f}(x) < \frac{k}{n}\log^{2d}(n/k)$. Indeed, the bias is large in the tail. In particular, it is possible that the kth nearest neighbour can be far from a test point, so that the $k$nn classifier is almost certain to classify differently to the Bayes classifier. Note that, for $d \leq \max\{4, \frac{4\rho}{\rho-4}\}$, we have $\rho \leq \max\{4, \frac{4d}{d-4}\}$; with too few moments the P_X-measure of $\{x \in \mathbb{R}^d : \bar{f}(x) < \frac{k}{n}\log^{2d}(n/k)\}$ can be large. In this case $\rho/(\rho+d) < 4/d$, and, without further assumptions, it may be the case that the excess risk in the tail is dominating the excess risk from the main body of the distribution.

If the marginal distribution of X has infinitely many moments, then for $d \leq 4$, part (ii) has the following corollary:

Corollary 3.3. *Assume **(B.1)**, **(B.2)**, **(B.3)** and **(B.4)(ρ)**, for all $\rho > 0$. Suppose that $d \leq 4$. Then for each $\beta \in (0,1)$ and all $\gamma > 0$,*

$$R(\hat{C}_n^{k\mathrm{nn}}) - R(C^{\mathrm{Bayes}}) = \frac{B_1}{k} + o\Big(\frac{1}{k} + \Big(\frac{k}{n}\Big)^{1-\gamma}\Big) \quad (3.13)$$

as $n \to \infty$, uniformly for $k \in K_{\beta,0}$.

Under the conditions of Theorem 3.2, part (i), we can balance the dominant terms in the expansion in (3.10) by setting

$$k^* := \Big\lfloor \Big(\frac{dB_1}{4B_2}\Big)^{d/(d+4)} n^{4/(d+4)} \Big\rfloor. \quad (3.14)$$

Note that this is the same as the choice in (3.2), except that that the constants B_1 and B_2 have changed. Under the assumptions of part (i), the excess risk of the k^*-nearest neighbour classifier satisfies

$$R(\hat{C}_n^{k^*\mathrm{nn}}) - R(C^{\mathrm{Bayes}}) = (d+4)\Big(\frac{B_1}{4}\Big)^{4/(d+4)}\Big(\frac{B_2}{d}\Big)^{d/(d+4)} n^{-4/(d+4)}\{1+o(1)\}. \quad (3.15)$$

In the case of Theorem 3.2, part (ii), however, there is no direct trade-off. To minimise the risk in the worst case one would chose, up to logarithmic terms, $k = O(n^{\rho/(2\rho+d)})$, and in the case that X has infinitely many moments, $k = O(n^{1/2})$.

Consider the important practical problem of choosing k. One could try to estimate the constants B_1 and B_2 in (3.14) and use a 'plug-in' choice of k. However, as discussed

in Samworth (2012b), estimating these is hard. One needs to first estimate the set $\mathcal{S}$, and then compute integrals of unknown functions over this set.

In practice, choosing k via a data-driven method works well, for example, using N-fold cross validation. Theorem 3.2 suggests a way to re-scale the cross validated choice based on the rate of convergence of the risk. Indeed, for 5-fold cross validation, the risks are estimated with a training set 4/5ths of the original size, thus, if the assumptions of part (i) are satisfied, one should re-scale the best cross validated choice by $(5/4)^{4/(d+4)}$, and if the assumptions of part (ii) are satisfied, then the appropriate re-scaling is $(5/4)^{\rho/(2\rho+d)}$. Of course, the number of moments needs to be estimated here too.

3.4 Tail adaptive classification

The results of Section 3.3, suggest that one may be able to achieve a faster rate of convergence for the global excess risk in some cases by letting k depend on the location of the test point.

3.4.1 Oracle classifier

To aid a preliminary discussion, suppose that the marginal density $\bar{f}$ is known. For $\beta \in (0,1)$, let

$$k_{\mathrm{O}}(x) := \max\bigl[1, \min\bigl\{\bigl\lfloor B\bigl\{\bar{f}(x)n\bigr\}^{4/(d+4)}\bigr\rfloor, \lfloor n^{1-\beta}\rfloor\bigr\}\bigr], \tag{3.16}$$

where the subscript O refers to the fact that this is an oracle choice of the function k_{L}, since it depends on $\bar{f}$. This choice balances the first order terms in the asymptotic expansions of the local squared-bias and variance of $\hat{S}_n(x)$, which, under our assumptions, are $O\bigl([k(x)/\{n\bar{f}(x)\}]^{4/d}\bigr)$ and $O(1/k(x))$, respectively. We start by stating the global excess risk result:

Theorem 3.4. *Assume **(B.1)**, **(B.2)**, **(B.3)** and **(B.4)(ρ)**. Then for each $\beta \in (0,1)$ and each $B > 0$,*

(i) if $\rho > 4$,

$$R(\hat{C}_n^{k_{\mathrm{O}}\mathrm{nn}}) - R(C^{\mathrm{Bayes}}) = B_3 n^{-4/(d+4)}\{1 + o(1)\} \tag{3.17}$$

as $n \to \infty$, where

$$B_3 := \frac{1}{B}\int_{\mathcal{S}} \frac{\bar{f}(x_0)^{d/(d+4)}}{4\|\dot{\eta}(x_0)\|}\, d\mathrm{Vol}^{d-1}(x_0) + B^{4/d}\int_{\mathcal{S}} \frac{\bar{f}(x_0)^{d/(d+4)}}{\|\dot{\eta}(x_0)\|} a(x)^2\, d\mathrm{Vol}^{d-1}(x_0);$$

(ii) if $\rho \leq 4$, then

$$R(\hat{C}_n^{k_{\mathrm{O}}\mathrm{nn}}) - R(C^{\mathrm{Bayes}}) = o(n^{-\rho/(\rho+d)} \log^{2(d+4)\rho/(\rho+d)} n) \tag{3.18}$$

as $n \to \infty$.

Note here that, in comparison to Theorem 3.2, the condition on ρ in Theorem 3.4 does not depend on the dimension d, and the conditions imposed in Theorem 3.4 to achieve a $O(n^{-4/(d+4)})$ global rate of convergence are much weaker than those in Theorem 3.2. The proof of Theorem 3.4 is similar to Theorem 3.2 and can be summarised as follows: in part (i), the main contribution arises from a neighbourhood of the Bayes decision boundary $\mathcal{S}$ inside the set $\mathcal{R}'_n := \{x \in \mathbb{R}^d : \bar{f}(x) \geq \frac{\log^{2(d+4)} n}{n}\}$. Then, as before, the moment condition ensures that the constant B_3 is finite, and gives that the risk over the set $\mathbb{R}^d \setminus \mathcal{R}'_n$ is negligible. In part (ii), however, the dominant contribution may arise from the tail, giving the little o term in (3.18).

Consider minimising the constant B_3 in (3.17). A simple calculation shows that the minimum is achieved by setting

$$B = B^* := \Biggl(\frac{d \int_{\mathcal{S}} \frac{\bar{f}(x_0)^{d/(d+4)}}{4\|\dot{\eta}(x_0)\|} \, d\mathrm{Vol}^{d-1}(x_0)}{4 \int_{\mathcal{S}} \frac{\bar{f}(x_0)^{d/(d+4)}}{\|\dot{\eta}(x_0)\|} a(x_0)^2 \, d\mathrm{Vol}^{d-1}(x_0)} \Biggr)^{\frac{d}{d+4}}.$$

This yields that

$$B_3^* := (d+4) \Big(\int_{\mathcal{S}} \frac{\bar{f}(x_0)^{d/(d+4)}}{16\|\dot{\eta}(x_0)\|} \, d\mathrm{Vol}^{d-1}(x_0) \Big)^{\frac{4}{d+4}} \Big(\int_{\mathcal{S}} \frac{\bar{f}(x_0)^{d/(d+4)}}{d\|\dot{\eta}(x_0)\|} a(x)^2 \, d\mathrm{Vol}^{d-1}(x_0) \Big)^{\frac{d}{d+4}}.$$

It is informative to compare B_3^* with the optimal constant for the standard $k$nn classifier given in (3.15). Indeed, by Hölder's inequality, we have that

$$B_1^{4/(d+4)} B_2^{d/(d+4)} \geq \int_{\mathcal{S}} \frac{\bar{f}(x_0)^{d/(d+4)}}{4^{4/(d+4)} \|\dot{\eta}(x_0)\|} a(x_0)^{2d/(d+4)} \, d\mathrm{Vol}^{d-1}(x_0).$$

Therefore

$$\frac{B_3^*}{(d+4)\big(B_1/4\big)^{4/(d+4)} \big(B_2/d\big)^{d/(d+4)}} \leq \Big(\frac{\sup_{x_0 \in \mathcal{S}} a(x_0)^2}{\inf_{x_0 \in \mathcal{S}} a(x_0)^2} \Big)^{d/(d+4)}.$$

Thus, if the function a is constant on $\mathcal{S}$, then the oracle tail adaptive classifier will (asymptotically) perform at least as well as the standard $k$nn classifier (with optimal k). In practice, of course, $\bar{f}$ is unknown. In the semi-supervised setting we propose to estimate $\bar{f}$ using the unlabelled training set.

3.4.2 The semi-supervised nearest neighbour classifier

Suppose that $\hat{f}_m$ is an estimate of $\bar{f}$, based on the unlabelled training set $\mathcal{T}'_m$. Many different techniques are available, including kernel methods, nearest neighbours, histograms, and many more. Silverman (1986) provides a good practical introduction to the subject; see also Jones et al. (1996). More recently, shape-constrained methods have enjoyed popularity: see, for example, Walther (2009); Chen and Samworth (2013); Kim and Samworth (2014) and the references therein.

Now, for $\beta \in (0,1)$, let

$$k_{\mathrm{SS}}(x) := \max[1, \min\{\lfloor B\{\hat{f}_m(x)n\}^{4/(d+4)}\rfloor, \lfloor n^{1-\beta}\rfloor\}], \tag{3.19}$$

where $\hat{f}_m$ denotes the kernel density estimator of $\bar{f}$, given by

$$\hat{f}_m(x) = \hat{f}_{m,h}(x) := \frac{1}{mh^d}\sum_{j=1}^{m} K\Big(\frac{x - X_{n+j}}{h}\Big),$$

where $K(x)$ is a bounded, square-integrable kernel function, with finite second moment, and is of the form $Q(p(x))$, for a polynomial p and a function of bounded variation Q.

Let $\mathcal{S}'_n := \mathcal{S} \cap \mathcal{R}'_n$. We have the following:

Theorem 3.5. *Assume* ***(B.1)***, ***(B.2)***, ***(B.3)*** *and* ***(B.4)(ρ)***. *Suppose further that* $m = m_n \geq m_0 n^{2+d/2}$, *for some sufficiently large* $m_0 > 0$, *and* $h = h_m := Am^{-1/(d+4)}$, *for some* $A > 0$. *Then for each* $\beta \in (0,1)$ *and each* $B > 0$,

(i) if $\rho > 4$,

$$R(\hat{C}_n^{k_{\mathrm{SS}}\mathrm{nn}}) - R(C^{\mathrm{Bayes}}) = B_4 n^{-4/(d+4)}\{1 + o(1)\} \tag{3.20}$$

as $n \to \infty$, *where*

$$B_4 := \mathbb{E}\Big\{\frac{1}{B}\int_{\mathcal{S}'_n} \frac{\bar{f}(x_0)^{d/(d+4)}}{4\|\dot{\eta}(x_0)\|}\Big(\frac{\bar{f}(x_0)}{\hat{f}_m(x_0)}\Big)^{4/(d+4)} d\mathrm{Vol}^{d-1}(x_0) + B^{4/d}\int_{\mathcal{S}'_n} \frac{\bar{f}(x_0)^{d/(d+4)}}{\|\dot{\eta}(x_0)\|}\Big(\frac{\hat{f}_m(x_0)}{\bar{f}(x_0)}\Big)^{16/(d(d+4))} a(x_0)^2\, d\mathrm{Vol}^{d-1}(x_0)\Big\}.$$

(ii) if $\rho \leq 4$, *then*

$$R(\hat{C}_n^{k_{\mathrm{SS}}\mathrm{nn}}) - R(C^{\mathrm{Bayes}}) = o(n^{-\rho/(\rho+d)} \log^{2(d+4)\rho/(\rho+d)} n)$$

as $n \to \infty$.

The lower bound on m is used to control the sup-norm loss of $\hat{f}_m$ as estimate of $\bar{f}$ over $\mathcal{R}'_n$. In fact, a similar result holds for any multivariate density estimator $\tilde{f}_m = \tilde{f}_{m_n}$

satisfying

$$\mathbb{P}\Big(\|\tilde{f}_m - f\|_\infty \geq \frac{\log^{2(d+4)} n}{2n}\Big) = o(n^{-4/(d+4)})$$

as $n \to \infty$. This ensures that $c \leq \inf_{x \in \mathcal{R}'_n} \frac{\bar{f}(x)}{\tilde{f}_m(x)} \leq \sup_{x \in \mathcal{R}'_n} \frac{\bar{f}(x)}{\tilde{f}_m(x)} \leq C$, for some $0 < c \leq C$, with high probability.

3.5 Empirical analysis

We compare the tail adaptive k_{O}nn and k_{SS}nn classifiers, introduced in the previous section, with the standard $k$nn classifier. We investigate three settings that reflect the differences between the main results in Sections 3.3 and 3.4.

- Setting 1: P_1 is the distribution of d independent Laplace components; P_2 is the d-dimensional multivariate Normal distribution with mean $\mu = (1, \ldots, 1)^T$ and covariance $I_{d\times d}$.

- Setting 2: P_1 is the distribution of d independent t_5 components; P_2 is the distribution of d independent components, the first $\lfloor d/2 \rfloor$ have a t_5 distribution and the remaining $d - \lfloor d/2 \rfloor$ a $N(1, 1)$ distribution.

- Setting 3: P_1 is the distribution of d independent standard Cauchy components; P_2 is the distribution of d independent components, the first $\lfloor d/2 \rfloor$ standard Cauchy and the remaining $d - \lfloor d/2 \rfloor$ standard Laplace.

The corresponding marginal distribution P_X in Setting 1 has infinitely many moments. Therefore, by Theorem 3.2, for $d \leq 4$, the global risk of the optimal standard $k$nn classifier has, up to log terms, a $O(n^{-1/2})$ rate of convergence. However, for the tail adaptive versions of the classifier, we can expect $O(n^{-4/(d+4)})$ rates. If $d \geq 5$, all three classifiers have a $O(n^{-4/(d+4)})$ rate. In Setting 2, we have only up to (but not including) 5 finite moments. In this case, by Theorem 3.2, for $d < 20$, the optimal standard $k$nn classifier has (up to log terms) a $O(n^{-5/(d+5)})$ rate of convergence. Whereas, we can expect a $O(n^{-4/(d+4)})$ rate for optimal the tail adaptive classifiers. Finally, in Setting 3, there is no first moment and the dominant contribution to the excess risk is arising from the tail for the standard and adaptive $k$nn classifiers, the rates of convergence expected are (up to log terms) $O(n^{-1/(d+2)})$ and $O(n^{-1/(d+1)})$, respectively.

We use 5-fold cross validation to choose the tuning parameter for each classifier. More precisely, for the standard $k$nn classifier, we set $k = \hat{k} = \lfloor (5/4)^{4/(d+4)} \tilde{k} \rfloor$, where $\tilde{k}$ is the value of k (from an equally spaced sequence of length at most 40 of values between 1 and $\lfloor n/4 \rfloor$) that yields the smallest 5-fold cross-validation error. The rescaling $(5/4)^{4/(d+4)}$ is used since $\tilde{k}$ is chosen based on a sample of size $4n/5$. Note that we are

Table 3.1: *Misclassification rates for Settings 1, 2 and 3. In the final two columns we present the regret ratios given in* (3.21) *(with standard errors calculated via the delta method). Highlighted in bold are the cases were there is a significant (more than 1 standard error) difference between the standard and semi-supervised classifiers.*

d	Bayes risk	n	$\hat{k}$nn risk	$\hat{k}_{\mathrm{O}}$nn risk	$\hat{k}_{\mathrm{SS}}$nn risk	O RR	SS RR
Setting 1							
1	30.02	50	$35.62_{0.15}$	$35.57_{0.14}$	$\mathbf{35.43}_{0.15}$	$0.99_{0.021}$	$0.97_{0.022}$
		200	$31.79_{0.08}$	$31.34_{0.07}$	$\mathbf{31.31}_{0.07}$	$0.75_{0.031}$	$0.73_{0.030}$
		1000	$30.73_{0.05}$	$30.30_{0.05}$	$\mathbf{30.33}_{0.05}$	$0.39_{0.046}$	$0.44_{0.043}$
2	24.21	50	$30.00_{0.12}$	$30.01_{0.12}$	$29.95_{0.12}$	$1.00_{0.017}$	$0.99_{0.017}$
		200	$26.73_{0.06}$	$26.04_{0.06}$	$\mathbf{26.08}_{0.06}$	$0.73_{0.018}$	$0.74_{0.019}$
		1000	$25.72_{0.05}$	$24.96_{0.04}$	$\mathbf{25.02}_{0.04}$	$0.50_{0.019}$	$0.54_{0.018}$
5	13.17	50	$21.49_{0.10}$	$21.76_{0.10}$	$21.60_{0.10}$	$1.03_{0.009}$	$1.01_{0.009}$
		200	$17.90_{0.05}$	$17.48_{0.05}$	$\mathbf{17.42}_{0.05}$	$0.91_{0.008}$	$0.90_{0.008}$
		1000	$16.03_{0.04}$	$15.43_{0.04}$	$\mathbf{15.48}_{0.04}$	$0.79_{0.008}$	$0.81_{0.008}$
Setting 2							
1	31.16	50	$36.45_{0.14}$	$36.07_{0.14}$	$\mathbf{35.93}_{0.14}$	$0.93_{0.021}$	$0.90_{0.021}$
		200	$32.80_{0.08}$	$32.38_{0.07}$	$\mathbf{32.42}_{0.07}$	$0.74_{0.034}$	$0.77_{0.035}$
		1000	$31.58_{0.05}$	$31.37_{0.05}$	$\mathbf{31.37}_{0.05}$	$0.51_{0.067}$	$0.51_{0.068}$
2	31.15	50	$37.79_{0.13}$	$38.02_{0.12}$	$37.90_{0.12}$	$1.03_{0.015}$	$1.02_{0.015}$
		200	$33.58_{0.08}$	$33.63_{0.07}$	$33.54_{0.07}$	$1.02_{0.029}$	$0.98_{0.027}$
		1000	$31.81_{0.05}$	$31.81_{0.05}$	$31.80_{0.05}$	$1.00_{0.039}$	$0.98_{0.039}$
5	20.10	50	$\mathbf{28.70}_{0.13}$	$29.16_{0.12}$	$29.13_{0.11}$	$1.05_{0.011}$	$1.05_{0.011}$
		200	$\mathbf{23.63}_{0.06}$	$23.77_{0.06}$	$23.93_{0.06}$	$1.03_{0.014}$	$1.08_{0.015}$
		1000	$21.85_{0.04}$	$21.70_{0.04}$	$\mathbf{21.77}_{0.04}$	$0.92_{0.013}$	$0.95_{0.014}$
Setting 3							
1	41.95	50	$47.59_{0.09}$	$46.63_{0.10}$	$\mathbf{46.58}_{0.10}$	$0.83_{0.014}$	$0.82_{0.014}$
		200	$45.53_{0.08}$	$44.66_{0.08}$	$\mathbf{44.74}_{0.08}$	$0.76_{0.018}$	$0.78_{0.019}$
		1000	$43.41_{0.06}$	$42.80_{0.06}$	$\mathbf{42.84}_{0.05}$	$0.58_{0.030}$	$0.61_{0.029}$
2	41.96	50	$47.99_{0.07}$	$47.16_{0.09}$	$\mathbf{47.23}_{0.09}$	$0.86_{0.011}$	$0.88_{0.012}$
		200	$46.36_{0.07}$	$45.63_{0.07}$	$\mathbf{45.67}_{0.08}$	$0.83_{0.013}$	$0.84_{0.014}$
		1000	$44.09_{0.06}$	$43.73_{0.06}$	$\mathbf{43.64}_{0.06}$	$0.83_{0.022}$	$0.79_{0.021}$
5	32.00	50	$45.05_{0.10}$	$43.00_{0.10}$	$\mathbf{43.15}_{0.12}$	$0.84_{0.006}$	$0.85_{0.006}$
		200	$41.15_{0.08}$	$38.93_{0.07}$	$\mathbf{39.35}_{0.08}$	$0.76_{0.007}$	$0.80_{0.007}$
		1000	$36.93_{0.05}$	$35.05_{0.05}$	$\mathbf{36.23}_{0.05}$	$0.82_{0.008}$	$0.86_{0.008}$

choosing the $\hat{k}$ to minimise the error from the body of the distribution – in some cases this is a sub-optimal choice for the global risk.

For the oracle tail adaptive classifier, we set

$$\hat{k}_{\mathrm{O}}(x) := \max\Big[1, \min\big[\lfloor \hat{B}_{\mathrm{O}}\{\bar{f}(x)n/\|\bar{f}\|_\infty\}^{4/(d+4)} \rfloor, n/2\big]\Big].$$

Similarly, for the semi-supervised classifier, we set

$$\hat{k}_{\mathrm{SS}}(x) := \max\Big[1, \min\big[\lfloor \hat{B}_{\mathrm{SS}}\{\hat{f}_m(x)n/\|\hat{f}_m\|_\infty\}^{4/(d+4)}\rfloor, n/2\big]\Big],$$

where $\hat{f}_m$ is the multidimensional kernel density estimator, with a truncated Normal kernel and bandwidths chosen via the default method in the `R` package `ks` (Duong, 2015). The choice of $\hat{B}_{\mathrm{O}}$ and $\hat{B}_{\mathrm{SS}}$ is made as follows: Let n_j denote the training sample size used in the jth fold of the cross-validation procedure, then for each possible choice of B_{O}, we use $\hat{k}_{\mathrm{O}}(x) = \max[1, \min\{\lfloor B_{\mathrm{O}}\{\bar{f}(x)n_j/\|\bar{f}\|_\infty\}^{4/(d+4)}\rfloor, n_j/2\}]$. (A similar method is used to choose $\hat{B}_{\mathrm{SS}}$.) The choices of $\hat{B}_{\mathrm{O}}$ and $\hat{B}_{\mathrm{SS}}$, therefore, do not require rescaling. In our simulations, $\hat{B}_{\mathrm{O}}$ and $\hat{B}_{\mathrm{SS}}$ are chosen from a sequence of 40 equally spaced points between $n^{-4/(d+4)}$ and $n^{d/(d+4)}$. Moreover, for the semi-supervised classifier we estimate $\|\hat{f}_m\|_\infty$ by the maximum value attained on the unlabelled training set. Note that using the lower limits, namely when $\hat{B}_{\mathrm{O}}, \hat{B}_{\mathrm{SS}} = n^{-4/(d+4)}$, yield the standard 1-nearest neighbour classifier.

In each of the 3 settings above, we used three different values of d, a training set of size $n \in \{50, 200, 1000\}$, an unlabelled training set of size 1000, and a test set of size 1000. In Table 3.1, we present the sample mean and standard error (in subscript) of the risks for 1000 repetitions of each experiment. Further, we present estimates of the regret ratios, given by

$$\frac{R(\hat{C}_n^{\hat{k}_{\mathrm{O}}\mathrm{nn}}) - R(C^{\mathrm{Bayes}})}{R(\hat{C}_n^{\hat{k}\mathrm{nn}}) - R(C^{\mathrm{Bayes}})} \quad \text{and} \quad \frac{R(\hat{C}_n^{\hat{k}_{\mathrm{SS}}\mathrm{nn}}) - R(C^{\mathrm{Bayes}})}{R(\hat{C}_n^{\hat{k}\mathrm{nn}}) - R(C^{\mathrm{Bayes}})}, \tag{3.21}$$

for which the standard errors given are estimated via the delta method.

In Table 3.1 we see improvement in performance from the tail adaptive classifiers in 20 of the 27 experiments, comparable performance (the risk estimates are within 1 standard error) in 5 of the 27, and there are just 2 settings where the standard $k$nn classifier is best.

In Setting 1, there is an improvement in each dimension when $n \geq 200$. When $n = 50$ the classifiers perform comparatively (the estimated regret ratio is within one standard error of 1). Possibly the sample size is too small for the asymptotic theory used in constructing the tail adaptive classifier to hold. Note further that the improvement is greater for the smaller values of d. This agrees with the result in Theorem 3.4, namely that the excess risk of tail adaptive classifier converges slower for larger d. In Setting 2, we only see an improvement when $d = 1$. For $d = 2, 5$ the improvement expected in theory is small, e.g for $d = 5$ the optimal standard $k$nn classifier has a $O(n^{-1/2})$ rate of convergence and the optimal tail adaptive classifier has a $O(n^{-4/9})$ rate – it is possible that the sample sizes used here are insufficient to detect the theoretical improvement

in rate and the leading constant is driving the performance. In Setting 3, we see improvement in performance from the oracle tail adaptive classifiers in every situation. Whilst the dominant contribution to the excess risk for all classifiers is arising from the tail, the improvement in rate is large. Again the improvement is greater for lower dimensional data.

A further comparison to draw from these experiments is between the oracle and semi-supervised version of the tail adaptive classifiers. The performance is the same (the risk estimates are within one standard error) in all cases except in Settings 2 and 3 and when $d = 5$. In this case, the semi-supervised version performs worse than the oracle. This is likely due to the fact that density estimation is harder in higher dimensions, if one had a larger unlabelled training data set we could expect similar performance.

3.6 Appendix

We first require some further notation. Define the $n \times d$ matrices $X^n := (X_1 \dots X_n)$ and $x^n := (x_1 \dots x_n)$. Write

$$\hat{\mu}_n(x) = \hat{\mu}_n(x, x^n) := \mathbb{E}\{\hat{S}_n(x)|X^n = x^n\} = \frac{1}{k_{\mathrm{L}}(x)} \sum_{i=1}^{k_{\mathrm{L}}(x)} \eta(x_{(i)}),$$

and

$$\hat{\sigma}_n^2(x) = \hat{\sigma}_n^2(x, x^n) := \mathrm{Var}\{\hat{S}_n(x)|X^n = x^n\} = \frac{1}{k_{\mathrm{L}}(x)^2} \sum_{i=1}^{k_{\mathrm{L}}(x)} \eta(x_{(i)})\{1 - \eta(x_{(i)})\}.$$

Here we have used the fact that the ordered labels $Y_{(1)}, \dots, Y_{(n)}$ are independent given X^n, satisfying $\mathbb{P}(Y_{(i)} = 1|X^n) = \eta(X_{(i)})$. Since η takes values in $[0, 1]$ it is clear that $0 \leq \hat{\sigma}_n^2(x) \leq \frac{1}{4k_{\mathrm{L}}(x)}$ for all $x \in \mathbb{R}^d$. Further, write $\mu_n(x) := \mathbb{E}\{\hat{S}_n(x)\} = \frac{1}{k_{\mathrm{L}}(x)} \sum_{i=1}^{k_{\mathrm{L}}(x)} \mathbb{E}\eta(X_{(i)})$ for the unconditional expectation of $\hat{S}_n(x)$. Finally, we will write $p_\delta(x) := P_X\big(B_\delta(x)\big)$.

Proof of Proposition 3.1. Fix $x \in \mathcal{R}^c$, and let z denote the orthogonal projection of x onto the boundary of $\mathcal{R}$, denoted $\partial\mathcal{R}$, so that $(x - z)^T(w - z) \leq 0$ for all $w \in \mathcal{R}$. First suppose that $x \neq z$, and observe that if $w_0 \in \mathcal{R}_0$, then $w_0 + \frac{2\delta(x-z)}{\|x-z\|} \in \mathcal{R}$, so

$$0 \geq (x - z)^T \left(w_0 + \frac{2\delta(x - z)}{\|x - z\|} - z \right) = (x - z)^T(w_0 - z) + 2\delta\|x - z\|.$$

We deduce that if $w_0 \in \mathcal{R}_0$, then

$$\begin{aligned}\|x - w_0\|^2 &= \|x - z\|^2 - 2(x - z)^T(w_0 - z) + \|z - w_0\|^2 \\ &> \|x - z\|^2 + 4\delta\|x - z\| + 4\delta^2 = \big(\|x - z\|^2 + 2\delta\big)^2.\end{aligned}$$

The same conclusion also holds when $x = z$. On the other hand, if $y \in B_{2\delta}(z)$, then

$$\|x - y\| \leq \|x - z\| + 2\delta.$$

It follows that if $\|X_{(k)}(z) - z\| \leq 2\delta$, then none of the k nearest neighbours of x belong to $\mathcal{R}_0$. Now, from the proof of Lemma 8.4.3 of Dudley (1999), there exist $z_1, \ldots, z_N \in \partial\mathcal{R}$ and $C_d > 0$ such that $N \leq C_d\delta^{-(d-1)}$ and given any $z \in \partial\mathcal{R}$, we can find $j \in \{1, \ldots, N\}$ such that $\|z - z_j\| \leq \delta$. Note that if $\|X_{(k)}(z_j) - z_j\| \leq \delta$ for $j = 1, \ldots, N$, then $\|X_{(k)}(z) - z\| \leq 2\delta$ for all $z \in \partial\mathcal{R}$. Let

$$A_k := \bigcap_{i=1}^{k} \bigcap_{x \in \mathcal{R}^c} \big\{(x_1, \ldots, x_n) : x_{(i)}(x) \notin \mathcal{R}_0\big\},$$

let $p_* := \inf_{z \in \partial\mathcal{R}} P_X\big(B_\delta(z)\big)$, and $T_1 \sim \mathrm{Bin}(n, p_*)$. Then by Hoeffding's inequality, for every $M > 0$,

$$\begin{aligned}\mathbb{P}(X^n \in A_k^c) &\leq \mathbb{P}\bigg(\max_{j=1,\ldots,N} \|X_{(k)}(z_j) - z_j\| > \delta\bigg) \\ &\leq \frac{C_d}{\delta^{d-1}}\mathbb{P}(T_1 < k) \leq \frac{C_d}{\delta^{d-1}} \exp\bigg(-\frac{2(np_* - k)^2}{n}\bigg) = O(n^{-M}),\end{aligned}$$

uniformly for $k \in K_{\beta,0}$. Now, by assumption on η, if $x^n \in A_k$, then

$$\inf_{x \in \mathcal{R}^c} \hat{\mu}_n(x, x^n) = \inf_{x \in \mathcal{R}^c} \frac{1}{k} \sum_{i=1}^{k} \eta(x_{(i)}(x)) \geq 1/2 + \epsilon/2.$$

Thus, by a further application of Hoeffding's inequality, we have that for each $M > 0$,

$$\begin{aligned}&\sup_{k \in K_{\beta,0}} \sup_{x \in \mathcal{R}^c} \mathbb{P}\{\hat{S}_n(x) < 1/2, X^n \in A_k\} \\ &\qquad = \sup_{k \in K_{\beta,0}} \sup_{x \in \mathcal{R}^c} \int_{A_k} \mathbb{P}\{\hat{S}_n(x) < 1/2 | X^n = x^n\}\, dP_X^n(x^n) \\ &\qquad \leq \sup_{k \in K_{\beta,0}} \sup_{x \in \mathcal{R}^c} \int_{A_k} \exp\Big(-\frac{\{\hat{\mu}_n(x) - 1/2\}^2}{2\hat{\sigma}_n^2(x)}\Big)\, dP_X^n(x^n) \\ &\qquad \leq \sup_{k \in K_{\beta,0}} \sup_{x \in \mathcal{R}^c} \exp\Big(-\frac{k\epsilon^2}{2}\Big) = O(n^{-M}),\end{aligned}$$

since $4\hat{\sigma}_n^2(x) \leq \frac{1}{k}$ for all $k \in K_{\beta,0}$. We conclude that

$$\sup_{k \in K_{\beta,0}} \{R_{\mathcal{R}^c}(\hat{C}_n^{k\mathrm{nn}}) - R_{\mathcal{R}^c}(C^{\mathrm{Bayes}})\}$$
$$= \sup_{k \in K_{\beta,0}} 2 \int_{\mathcal{R}^c} \mathbb{P}\{\hat{S}_n(x) < 1/2\}\{\eta(x) - 1/2\}\, dP_X(x) = O(n^{-M}),$$

for each $M > 0$, as $n \to \infty$. □

3.6.1 Asymptotic expansions

We prove the main results in Section 3.2.1. First, we derive an asymptotic expansion for the excess risk of the local-k-nearest neighbour (k_{L}nn) classifier over the region where $\bar{f}(x) \geq \frac{k_{\mathrm{L}}(x)}{n} \log^{2d}(n/k_{\mathrm{L}}(x))$. For a point x in this region, its $k_{\mathrm{L}}(x)$ nearest neighbours will concentrate on a shrinking ball at x. We can therefore derive asymptotic expansions for the bias and variance of $\hat{S}_n(x)$, and using the Berry–Esseen Theorem to approximate the probability of error, derive an asymptotic expansion for the excess risk of the k_{L}nn classifier, uniformly for k_{L} in a certain class of functions. Let $\mathcal{R}_n := \big\{x \in \mathbb{R}^d : \bar{f}(x) \geq \frac{k_{\mathrm{L}}(x)}{n} \log^{2d}\big(\frac{n}{k_{\mathrm{L}}(x)}\big)\big\}$, and note that $\mathcal{R}_n$ is compact, since Assumption **(B.3)** ensures that $\bar{f}(x) \to 0$ as $\|x\| \to \infty$. Recall that $\mathcal{S} = \{x \in \mathbb{R}^d : \eta(x) = 1/2\}$. For $\beta \in (0,1)$, define the class of functions

$$K_\beta := \{k_{\mathrm{L}} : \mathbb{R}^d \to \{\lceil \log^4 n \rceil, \ldots, \lfloor n^{1-\beta} \rfloor\} | k_{\mathrm{L}}(x) = \lfloor g(x) \rfloor,\ g : \mathbb{R}^d \to \mathbb{R} \text{ continuous}\}.$$

Theorem 3.6. *Assume **(B.1)**, **(B.2)**, **(B.3)** and **(B.4)(ρ)**, for some $\rho > 0$. Then for each $\beta \in (0,1)$,*

$$R_{\mathcal{R}_n}(\hat{C}_n^{k_{\mathrm{L}}\mathrm{nn}}) - R_{\mathcal{R}_n}(C^{\mathrm{Bayes}}) = \gamma_n(k_{\mathrm{L}})\{1 + o(1)\} + o\Big(\sup_{x \in \mathcal{R}_n^c} \bar{f}(x)\Big)$$

as $n \to \infty$, uniformly for $k_{\mathrm{L}} \in K_\beta$, where

$$\gamma_n(k_{\mathrm{L}}) := \int_{S \cap \mathcal{R}_n} \frac{\bar{f}(x_0)}{4k_{\mathrm{L}}(x_0)\|\dot{\eta}(x_0)\|}\, d\mathrm{Vol}^{d-1}(x_0)$$
$$+ \frac{1}{n^{4/d}} \int_{S \cap \mathcal{R}_n} \frac{\bar{f}(x_0)^{1-4/d} k_{\mathrm{L}}(x_0)^{4/d}}{\|\dot{\eta}(x_0)\|} a(x_0)^2\, d\mathrm{Vol}^{d-1}(x_0).$$

Proof of Theorem 3.6. First observe that

$$R_{\mathcal{R}_n}(\hat{C}_n^{k_{\mathrm{L}}\mathrm{nn}}) - R_{\mathcal{R}_n}(C^{\mathrm{Bayes}})$$
$$= \int_{\mathcal{R}_n} \big[\mathbb{P}\{\hat{S}_n(x) < 1/2\} - \mathbb{1}_{\{\eta(x) < 1/2\}}\big]\{2\eta(x) - 1\}\bar{f}(x)\, dx. \qquad (3.22)$$

Let $\mathcal{S}_n := \mathcal{S} \cap \mathcal{R}_n$. For $\epsilon > 0$, let

$$\mathcal{S}_n^{\epsilon\epsilon} := \{x \in \mathbb{R}^d : \eta(x) = 1/2 \text{ and } \mathrm{dist}(x, \mathcal{S}_n) < \epsilon\},$$

where $\mathrm{dist}(x, \mathcal{S}_n) := \inf_{x_0 \in \mathcal{S}_n} \|x - x_0\|$ and

$$\mathcal{S}_n^{\epsilon} := \Big\{x_0^t := x_0 + t\frac{\dot{\eta}(x_0)}{\|\dot{\eta}(x_0)\|} : x_0 \in \mathcal{S}_n^{\epsilon\epsilon}, |t| < \epsilon\Big\}.$$

The proof is presented in seven steps. We will see that the dominant contribution to the integral in (3.22) arises for a small neighbourhood about the Bayes decision boundary, i.e. the region $\mathcal{S}_n^{\epsilon_n} \cap \mathcal{R}_n$, where $\epsilon_n := \frac{1}{\log^{1/2}(n^\beta)}$. Outside of this region, the k_{L}nn classifier agrees with the Bayes classifier with high probability (asymptotically). More precisely, we show in Step 4 that

$$\sup_{k_{\mathrm{L}} \in K_\beta} \sup_{x \in \mathcal{R}_n \setminus \mathcal{S}_n^{\epsilon_n}} |\mathbb{P}\{\hat{S}_n(x) < 1/2\} - \mathbb{1}_{\{\eta(x) < 1/2\}}| = O(n^{-M}),$$

for each $M > 0$, as $n \to \infty$. In Steps 1, 2 and 3, we derive asymptotic expansions for the bias, conditional (on X^n) bias and variance of $\hat{S}_n(x)$. In Step 5 we show that the integral over $\mathcal{S}_n^{\epsilon}$ can be decomposed to one over $\mathcal{S}_n$ and one perpendicular to $\mathcal{S}_n$. Step 6 is dedicated to combining the results of Steps 1 - 5; we derive the leading order terms in the asymptotic expansion of the integral in (3.22). Finally, we bound the remaining error terms to conclude the proof in Step 7. To ease notation, in Steps 1, 2, 3 and 4 we write k_{L} in place of $k_{\mathrm{L}}(x)$.

Step 1: Let $\mu_n(x) := \mathbb{E}\{\hat{S}_n(x)\}$. We show that

$$\mu_n(x) - \eta(x) - \Big(\frac{k_{\mathrm{L}}}{n\bar{f}(x)}\Big)^{2/d} a(x) = o\Big(\Big(\frac{k_{\mathrm{L}}}{n\bar{f}(x)}\Big)^{2/d}\{1 + |a(x)|\}\Big),$$

uniformly for $k_{\mathrm{L}} \in K_\beta$ and $x \in \mathcal{S}_n^{\epsilon_n} \cap \mathcal{R}_n$. Write

$$\begin{aligned}\mu_n(x) &= \eta(x) + \frac{1}{k_{\mathrm{L}}}\sum_{i=1}^{k_{\mathrm{L}}} \mathbb{E}\{\eta(X_{(i)}) - \eta(x)\} \\ &= \eta(x) + \frac{1}{k_{\mathrm{L}}}\sum_{i=1}^{k_{\mathrm{L}}} \mathbb{E}\{(X_{(i)} - x)^T\dot{\eta}(x)\} + \frac{1}{2}\mathbb{E}\{(X_{(i)} - x)^T\ddot{\eta}(x)(X_{(i)} - x)\} + R_1,\end{aligned}$$

where we show in Step 7 that $|R_1| = o\big(\{k_{\mathrm{L}}/n\bar{f}(x)\}^{2/d}\big)$, uniformly for $x \in \mathcal{S}_n^{\epsilon_n} \cap \mathcal{R}_n$.

Now, the density of $X_{(i)} - x$ at $u \in \mathbb{R}^d$ is given by

$$f_{(i)}(u) := n\bar{f}(x+u)\binom{n-1}{i-1} p_{\|u\|}^{i-1}(1 - p_{\|u\|})^{n-i} = n\bar{f}(x+u)p_{\|u\|}^{n-1}(i-1),$$

where $p_{\|u\|}^{n-1}(i-1)$ denotes the probability that a $\mathrm{Bin}(n-1, p_{\|u\|})$ random variable equals $i-1$. Let $\delta_n = \delta_n(x) := \big(\frac{2k_{\mathrm{L}}}{n\bar{f}(x)a_d}\big)^{1/d}$. We show in Step 7 that

$$R_2 := \sup_{k_{\mathrm{L}} \in K_\beta} \sup_{x \in \mathcal{R}_n} \mathbb{E}\{\|X_{(k_{\mathrm{L}})} - x\|^2 \mathbb{1}_{\{\|X_{(k_{\mathrm{L}})} - x\| > \delta_n\}}\} = O(n^{-M}), \tag{3.23}$$

for each $M > 0$, as $n \to \infty$. It follows that

$$\mathbb{E}\{(X_{(i)} - x)^T \dot{\eta}(x)\} = \int_{B_{\delta_n}(0)} \dot{\eta}(x)^T u n \{\bar{f}(x+u) - \bar{f}(x)\} p_{\|u\|}^{n-1}(i-1)\, du + O(n^{-M}),$$

uniformly for $1 \le i \le k_{\mathrm{L}}$ and $x \in \mathcal{R}_n$. Moreover

$$\mathbb{E}\{(X_{(i)} - x)^T \ddot{\eta}(x)(X_{(i)} - x)\} = \int_{B_{\delta_n}(0)} u^T \ddot{\eta}(x) u n \bar{f}(x+u) p_{\|u\|}^{n-1}(i-1)\, du + O(n^{-M}),$$

uniformly for $1 \le i \le k_{\mathrm{L}}$ and $x \in \mathcal{R}_n$. Hence, summing over i, we see that

$$\begin{aligned}
&\frac{1}{k_{\mathrm{L}}} \sum_{i=1}^{k_{\mathrm{L}}} \mathbb{E}\{(X_{(i)} - x)^T \dot{\eta}(x)\} + \frac{1}{2k_{\mathrm{L}}} \sum_{i=1}^{k_{\mathrm{L}}} \mathbb{E}\{(X_{(i)} - x)^T \ddot{\eta}(x)(X_{(i)} - x)\} \\
&= \int_{B_{\delta_n}(0)} \Big[\dot{\eta}(x)^T u n \{\bar{f}(x+u) - \bar{f}(x)\} + \frac{1}{2} u^T \ddot{\eta}(x) u n \bar{f}(x+u)\Big] q_{\|u\|}^{n-1}(k_{\mathrm{L}})\, du + O(n^{-M}),
\end{aligned}$$

where $q_{\|u\|}^{n-1}(k_{\mathrm{L}})$ denotes the probability that a $\mathrm{Bin}(n-1, p_{\|u\|})$ random variable is less than k_{L}. By a Taylor expansion of $\bar{f}$ and assumption **(B.3)**, there exists $C_0 > 0$ such that

$$\Big|\bar{f}(x+t) - \bar{f}(x) - t^T \dot{\bar{f}}(x)\Big| \le \frac{\|t\|^2}{2} \sup_{u \in B_{\|t\|}(0)} \|\ddot{\bar{f}}(x+u)\|_{\mathrm{op}} \le C_0 \|t\|^2 \bar{f}(x) \log^2\Big(\frac{\bar{f}(x)}{2\|\bar{f}\|_\infty}\Big),$$

for all n sufficiently large, all $x \in \mathcal{R}_n$, and all $\|t\| \le \delta_n$. Hence, for all n sufficiently large,

$$\begin{aligned}
|p_{\|u\|} - \bar{f}(x) a_d \|u\|^d| &\le \int_{\|t\| \le \|u\|} |\bar{f}(x+t) - \bar{f}(x) - t^T \dot{\bar{f}}(x)|\, dt \\
&\le C_0 \bar{f}(x) \log^2\{\bar{f}(x)/(2\|\bar{f}\|_\infty)\} \int_{\|t\| \le \|u\|} \|t\|^2\, dt \\
&\le C_0 \bar{f}(x) \log^2\{\bar{f}(x)/(2\|\bar{f}\|_\infty)\} \|u\|^{d+2} \frac{d a_d}{d+2}
\end{aligned} \tag{3.24}$$

for all $\|u\| \le \delta_n$.

Let $b_n = b_n(k_{\mathrm{L}}) := \big(\frac{(n-1)\bar{f}(x)a_d}{k_{\mathrm{L}}}\big)^{1/d}$. By (3.24), we have that

$$k_{\mathrm{L}} - (n-1) p_{\|v\|/b_n} = k_{\mathrm{L}} - (n-1)\bar{f}(x) a_d \|v\|^d / b_n^d + R_4 = k_{\mathrm{L}}(1 - \|v\|^d) + R_4,$$

where

$$|R_4| \leq \frac{(n-1)C_0\bar{f}(x)\log\{\bar{f}(x)/(2\|\bar{f}\|_\infty)\}\|v\|^{d+2}da_d}{(d+2)b_n^{d+2}}$$
$$= \frac{k_{\mathrm{L}}C_0\log^2\{\bar{f}(x)/(2\|\bar{f}\|_\infty)\}\|v\|^{d+2}d}{(d+2)b_n^2} \leq \frac{k_{\mathrm{L}}C_0 2^{d+2}d}{a_d^{2/d}(d+2)\log^2(n/k_{\mathrm{L}})},$$

since $b_n^2 \geq a_d^{2/d}\log^4(n/k_{\mathrm{L}})$, for $x \in \mathcal{R}_n$. It follows that there exists $n_0 \in \mathbb{N}$ such that, for all $x \in \mathcal{R}_n$ and all $\|v\|^d \in (0, 1-2/\log(n/k_{\mathrm{L}})]$,

$$k_{\mathrm{L}} - (n-1)p_{\|v\|/b_n} \geq \frac{k_{\mathrm{L}}}{\log(n/k_{\mathrm{L}})}, \tag{3.25}$$

for all $n > n_0$. Similarly, for all $\|v\|^d \in [1+2/\log(n/k_{\mathrm{L}}), 2]$,

$$(n-1)p_{\|v\|/b_n} - k_{\mathrm{L}} \geq \frac{k_{\mathrm{L}}}{\log(n/k_{\mathrm{L}})}.$$

Hence, by Bernstein's inequality, we have that

$$\sup_{k_{\mathrm{L}}\in K_\beta}\sup_{x\in\mathcal{R}_n}\sup_{\|v\|^d\in(0,1-2/\log(n/k_{\mathrm{L}})]} 1 - q_{\|v\|/b_n}^{n-1}(k_{\mathrm{L}}) \leq \exp\Big(-\frac{\log^2 n}{\beta^2}\Big) = O(n^{-M}), \tag{3.26}$$

and

$$\sup_{k_{\mathrm{L}}\in K_\beta}\sup_{x\in\mathcal{R}_n}\sup_{\|v\|^d\in[1+2/\log(n/k_{\mathrm{L}}),2]} q_{\|v\|/b_n}^{n-1}(k_{\mathrm{L}}) \leq \exp\Big(-\frac{\log^2 n}{\beta^2}\Big) = O(n^{-M}), \tag{3.27}$$

for each $M > 0$, as $n \to \infty$. We conclude that

$$\frac{1}{k_{\mathrm{L}}}\int_{B_{\delta_n}(0)}[\dot{\eta}(x)^T un\{\bar{f}(x+u) - \bar{f}(x)\} + \frac{1}{2}u^T\ddot{\eta}(x)un\bar{f}(x+u)]q_{\|u\|}^{n-1}(k_{\mathrm{L}})\,du$$
$$= \frac{1}{k_{\mathrm{L}}}\int_{\|u\|\leq 1/b_n}\Big[\dot{\eta}(x)^T un\{\bar{f}(x+u) - \bar{f}(x)\} + \frac{1}{2}u^T\ddot{\eta}(x)un\bar{f}(x+u)\Big]du\{1+o(1)\}$$
$$= \Big(\frac{k_{\mathrm{L}}}{n}\Big)^{2/d}\frac{\sum_{j=1}^d\{\eta_j(x)\bar{f}_j(x) + \frac{1}{2}\eta_{jj}(x)\bar{f}(x)\}}{(d+2)a_d^{2/d}\bar{f}(x)^{1+2/d}}\{1+o(1)\}$$
$$= \Big(\frac{k_{\mathrm{L}}}{n\bar{f}(x)}\Big)^{2/d}a(x)\{1+o(1)\} \tag{3.28}$$

as $n \to \infty$, uniformly for $x \in \mathcal{R}_n$ and $k_{\mathrm{L}} \in K_\beta$. Here we have used the fact that $\int_{B_1(0)} v_j^2 dv = \frac{a_d}{d+2}$, for $j = 1, \ldots, d$.

Step 2: Recall that $\hat{\sigma}_n^2(x, x^n) = \mathrm{Var}\{\hat{S}_n(x)|X^n = x^n\}$. We show that

$$\Big|\hat{\sigma}_n^2(x, X^n) - \frac{1}{4k_{\mathrm{L}}}\Big| = o_p(1/k_{\mathrm{L}}), \tag{3.29}$$

uniformly for $k_{\mathrm{L}} \in K_\beta$ and $x \in \mathcal{S}_n^{\epsilon_n} \cap \mathcal{R}_n$.

To show (3.29): write

$$\hat{\sigma}_n^2(x, X^n) = \frac{1}{k_{\mathrm{L}}^2} \sum_{i=1}^{k_{\mathrm{L}}} \eta(X_{(i)})\{1 - \eta(X_{(i)})\} = \frac{1}{k_{\mathrm{L}}^2} \sum_{i=1}^{k_{\mathrm{L}}} \eta(X_{(i)}) - \frac{1}{k_{\mathrm{L}}^2} \sum_{i=1}^{k_{\mathrm{L}}} \eta(X_{(i)})^2.$$

Define $A_k := \{x^n : \|x_{(k)} - x\| < \epsilon_n/2, \text{for all } x \in \mathcal{R}_n\}$. Observe, by a similar argument to that leading to (3.25), that $p_\epsilon(x) > \frac{a_d}{2}\epsilon^d \bar{f}(x) \geq \frac{a_d}{2}\epsilon^d \frac{k_{\mathrm{L}}}{n} \log^{2d}(n/k_{\mathrm{L}})$, for all $\epsilon < \epsilon_n$ and all $x \in \mathcal{R}_n$. Now suppose that $z_1, \ldots, z_N \in \mathcal{R}_n$ are such that $\|z_i - z_j\| > \epsilon_n/3$, but $\sup_{x \in \mathcal{R}_n} \min_{i=1,\ldots,N} \|x - z_i\| \leq \epsilon_n/3$. We have that

$$1 = P_X(\mathbb{R}^d) \geq \sum_{i=1}^{N} p_{\epsilon_n/6}(z_i) \geq \frac{a_d N}{2} (\epsilon_n/6)^d \frac{\log^4(n)}{n} \log^{2d}(n^\beta).$$

It follows that

$$N \leq \frac{2^{d+1} 3^d n}{a_d \log^4(n) \log^{3d/2}(n^\beta)}.$$

Now, given $x \in \mathcal{R}_n$, let $i_0 := \operatorname{argmin} \|x - z_i\|$, so that $B_{\epsilon_n/6}(z_{i_0}) \subseteq B_{\epsilon_n/2}(x)$. Thus, if there are k points inside each of the balls $B_{\epsilon_n/6}(z_i)$, then there are k inside the ball of radius $\epsilon_n/2$ about every $x \in \mathcal{R}_n$. Moreover, there exists $\kappa > 0$, such that

$$\begin{aligned} np_{\epsilon_n/6}(x) - k_{\mathrm{L}} &\geq n\kappa \Big(\frac{\epsilon_n}{6}\Big)^d \bar{f}(x) - k_{\mathrm{L}} \geq n\kappa \Big(\frac{\epsilon_n}{6}\Big)^d \frac{k_{\mathrm{L}}}{n} \log^{2d}(n/k_{\mathcal{L}}) - k_{\mathrm{L}} \\ &\geq k_{\mathrm{L}}\{\kappa \log^{3d/2}(n^\beta) - 1\} \geq \log^4(n)\{\kappa \log^{3d/2}(n^\beta) - 1\}, \end{aligned}$$

for all $k_{\mathrm{L}} \in K_\beta$ and $x \in \mathcal{R}_n$. Hence, by Bernstein's inequality, we conclude that

$$\begin{aligned} 1 - \mathbb{P}(X^n \in A_k) &= \mathbb{P}\Big\{ \sup_{x \in \mathcal{R}_n} \|X_{(k)} - x\| \geq \epsilon_n/2 \Big\} \\ &\leq \mathbb{P}\Big\{ \max_{i=1,\ldots,N} \|X_{(k)}(z_i) - z_i\| \geq \epsilon_n/6 \Big\} \\ &\leq \sum_{i=1}^{N} \mathbb{P}\big\{ \|X_{(k)}(z_i) - z_i\| \geq \epsilon_n/6 \big\} \leq N \exp(-\log^4(n)/2) = O(n^{-M}), \end{aligned}$$

uniformly for $k \in [\log^4 n, n^{1-\beta}]$.

Now, we have that

$$\sup_{k_{\mathrm{L}} \in K_\beta} \sup_{x \in \mathcal{R}_n \cap \mathcal{S}^{\epsilon_n}} \sup_{x^n \in A_{k_{\mathrm{L}}}} \sup_{1 \leq i \leq k_{\mathrm{L}}(x)} |\eta(x_{(i)}) - 1/2| \to 0.$$

It follows that

$$\sup_{x\in\mathcal{R}_n\cap\mathcal{S}_n^{\epsilon_n}}\sup_{x^n\in A_{k_{\mathrm{L}}}}\Big|\frac{1}{k_{\mathrm{L}}^2}\sum_{i=1}^{k_{\mathrm{L}}}\{\eta(x_{(i)})-1/2\}\Big|=o\Big(\frac{1}{k_{\mathrm{L}}}\Big)$$

as $n\to\infty$, uniformly for $k_{\mathrm{L}}\in K_\beta$. Similarly,

$$\sup_{x\in\mathcal{R}_n\cap\mathcal{S}^{\epsilon_n}}\sup_{x^n\in A_{k_{\mathrm{L}}}}\Big|\frac{1}{k_{\mathrm{L}}^2}\sum_{i=1}^{k_{\mathrm{L}}}\{\eta(x_{(i)})^2-1/4\}\Big|=o\Big(\frac{1}{k_{\mathrm{L}}}\Big)$$

as $n\to\infty$, uniformly for $k_{\mathrm{L}}\in K_\beta$, and we conclude (3.29).

Step 3: Recall that $\hat{\mu}_n(x,x^n)=\mathbb{E}\{\hat{S}_n(x)|X^n=x^n\}$. We show that

$$\mathrm{Var}\{\hat{\mu}_n(x,X^n)\}=o\Big(\frac{1}{k_{\mathrm{L}}}+\Big(\frac{k_{\mathrm{L}}}{n\bar{f}(x)}\Big)^{4/d}a(x)^2\Big) \tag{3.30}$$

as $n\to\infty$, uniformly for $x\in\mathcal{R}_n$ and $k_{\mathrm{L}}\in K_\beta$. By writing the variance of the sum as the sum of covariance terms, observe first that

$$\begin{aligned}\mathrm{Var}\{\hat{\mu}_n(x,X^n)\}=&\frac{1}{k_{\mathrm{L}}^2}\sum_{i=1}^{k_{\mathrm{L}}}\mathbb{E}\{\eta(X_{(i)})^2\}-\mathbb{E}\{\eta(X_{(i)})\}^2\\&+\frac{2}{k_{\mathrm{L}}^2}\sum_{j=2}^{k_{\mathrm{L}}}\sum_{i=1}^{j-1}\mathbb{E}\{\eta(X_{(i)})\eta(X_{(j)})\}-\mathbb{E}\{\eta(X_{(i)})\}\mathbb{E}\{\eta(X_{(j)})\}.\end{aligned} \tag{3.31}$$

Then, we have that

$$\begin{aligned}\frac{1}{k_{\mathrm{L}}^2}\Big|\sum_{i=1}^{k_{\mathrm{L}}}\{\mathbb{E}\eta(X_{(i)})^2-\eta(x)^2\}\Big|&\le\frac{1}{k_{\mathrm{L}}}\sup_{1\le i\le k_{\mathrm{L}}}|\mathbb{E}\{\eta(X_{(i)})^2\}-\eta(x)^2|\\&\le\frac{2}{k_{\mathrm{L}}}\sup_{1\le i\le k_{\mathrm{L}}}\mathbb{E}|\eta(X_{(i)})-\eta(x)|=o\Big(\frac{1}{k_{\mathrm{L}}}\Big).\end{aligned}$$

Similarly

$$\frac{1}{k_{\mathrm{L}}^2}\Big|\sum_{i=1}^{k_{\mathrm{L}}}\{\mathbb{E}\eta(X_{(i)})\}^2-\eta(x)^2\Big|\le\frac{2}{k_{\mathrm{L}}}\sup_{1\le i\le k_{\mathrm{L}}}\mathbb{E}|\eta(X_{(i)})-\eta(x)|=o\Big(\frac{1}{k_{\mathrm{L}}}\Big),$$

uniformly for $x\in\mathcal{R}_n$ and $k_{\mathrm{L}}\in K_\beta$.

Now, for the second term in (3.31), note that for $x\in\mathbb{R}^d$ and $i<j$ the joint density of $(X_{(i)}-x,X_{(j)}-x)$ at (u_1,u_2), with $\|u_1\|<\|u_2\|$, is given by

$$f_{i,j}(u_1,u_2):=\bar{f}(x+u_1)\bar{f}(x+u_2)n(n-1)\mathbb{P}\{(I,J,K)=(i-1,j-i-1,n-j)\}.$$

where $(I,J,K)\sim\mathrm{Multi}(n-2,p_{\|u_1\|},p_{\|u_2\|}-p_{\|u_1\|},1-p_{\|u_2\|})$. By writing the multinomial

probability as a product of two binomial probabilities, we see that

$$f_{i,j}(u_1,u_2) = \bar{f}(x+u_1)\bar{f}(x+u_2)n(n-1)r^{j-2}_{\|u_1\|,\|u_2\|}(i-1)p^{n-2}_{\|u_2\|}(j-2),$$

where $r^{j-2}_{\|u_1\|,\|u_2\|}(i-1)$ denotes the probability that a $\mathrm{Bin}\big(j-2, \frac{p_{\|u_1\|}}{p_{\|u_2\|}}\big)$ random variable equals $i-1$. To ease notation let

$$\gamma(x,u_1,u_2) := \{\eta(x+u_1)-\eta(x)\}\{\eta(x+u_2)-\eta(x)\}\bar{f}(x+u_1)\bar{f}(x+u_2). \tag{3.32}$$

Observe that, for $x \in \mathcal{R}_n$, $2 \le j \le k_{\mathrm{L}}$ and n sufficiently large, we have that

$$\begin{aligned}
&\frac{2}{k_{\mathrm{L}}^2}\sum_{j=2}^{k_{\mathrm{L}}}\sum_{i=1}^{j-1}\mathbb{E}\big[\{\eta(X_{(i)})-\eta(x)\}\{\eta(X_{(j)})-\eta(x)\}\big] \\
&\quad= \frac{2n(n-1)}{k_{\mathrm{L}}^2}\sum_{j=2}^{k_{\mathrm{L}}}\int_{\mathbb{R}^d}\int_{u_1:\|u_1\|<\|u_2\|}\gamma(x,u_1,u_2)p^{n-2}_{\|u_2\|}(j-2)\,du_1\,du_2 \\
&\quad= \frac{2n(n-1)}{k_{\mathrm{L}}^2}\int_{\mathbb{R}^d}\int_{u_1:\|u_1\|<\|u_2\|}\gamma(x,u_1,u_2)q^{n-2}_{\|u_2\|}(k_{\mathrm{L}}-1)\,du_1\,du_2
\end{aligned}$$

since, for all $\|u_1\| \le \|u_2\|$, $\sum_{i=1}^{j-1} r^{j-2}_{\|u_1\|,\|u_2\|}(i-1) = 1$. Let $b'_n := b_{n-1}(k_{\mathrm{L}}-1) = \big(\frac{(n-2)\bar{f}(x)a_d}{k_{\mathrm{L}}-1}\big)^{1/d}$. By similar arguments to those in (3.26) and (3.27), that lead to (3.28), we have that

$$\begin{aligned}
&\frac{2n(n-1)}{k_{\mathrm{L}}^2}\int_{\mathbb{R}^d}\int_{u_1:\|u_1\|<\|u_2\|}\gamma(x,u_1,u_2)q^{n-2}_{\|u_2\|}(k_{\mathrm{L}}-1)\,du_1\,du_2 \\
&= \frac{2n(n-1)}{k_{\mathrm{L}}^2}\int_{\mathbb{R}^d}\int_{\|u_1\|<1}\|u_2\|^d\gamma(x,\|u_2\|u_1,u_2)q^{n-2}_{\|u_2\|}(k_{\mathrm{L}}-1)\,du_1\,du_2 \\
&= \frac{2n(n-1)}{k_{\mathrm{L}}^2}\int_{\|u_2\|\le 1/b'_n}\int_{\|u_1\|<1}\|u_2\|^d\gamma(x,u_1\|u_2\|,u_2)\,du_1\,du_2\{1+o(1)\} \\
&= \frac{2n(n-1)}{b_n'^{2d}k_{\mathrm{L}}^2}\int_{\|u_2\|\le 1}\int_{\|u_1\|<1}\|u_2\|^d\gamma(x,u_1\|u_2\|/b'_n,u_2b'_n)\,du_1\,du_2\{1+o(1)\} \\
&= \frac{2n(n-1)(k_{\mathrm{L}}-1)^{2+4/d}}{(n-2)^{2+4/d}k_{\mathrm{L}}^2}\frac{a(x)^2C_d}{\bar{f}(x)^{4/d}c_d}\{1+o(1)\} = \frac{2C_d}{c_d}\Big(\frac{k_{\mathrm{L}}}{n\bar{f}(x)}\Big)^{4/d}a(x)^2\{1+o(1)\},
\end{aligned} \tag{3.33}$$

where $c_d := \int_{\|t\|\le 1} t_1^2\,dt$ and $C_d := \int_{\|t\|\le 1}\|t\|^{d+2}t_1^2\,dt$. Now, for the sum of terms involving

$\mathbb{E}\{\eta(X_{(i)}) - \eta(x)\}\mathbb{E}\{\eta(X_{(j)}) - \eta(x)\}$, we have that

$$\begin{aligned}
&\frac{1}{k_{\mathrm{L}}^2}\sum_{j=2}^{k_{\mathrm{L}}}\sum_{i=1}^{j-1}\mathbb{E}\{\eta(X_{(i)}) - \eta(x)\}\mathbb{E}\{\eta(X_{(j)}) - \eta(x)\} \\
&\quad= \frac{n^2}{k_{\mathrm{L}}^2}\sum_{j=2}^{k_{\mathrm{L}}}\sum_{i=1}^{j-1}\int_{u_2\in\mathbb{R}^d}\int_{u_1\in\mathbb{R}^d}\gamma(x,u_1,u_2)p_{\|u_2\|}^{n-1}(j-1)p_{\|u_1\|}^{n-1}(i-1)\,du_1\,du_2 \\
&\quad= \frac{n^2}{k_{\mathrm{L}}^2}\sum_{j=2}^{k_{\mathrm{L}}}\int_{u_2\in\mathbb{R}^d}\int_{u_1\in\mathbb{R}^d}\gamma(x,u_1,u_2)p_{\|u_2\|}^{n-1}(j-1)q_{\|u_1\|}^{n-1}(j-1)\,du_1\,du_2.
\end{aligned}$$

By relabelling the indices in the sum, it follows that

$$\begin{aligned}
&\frac{1}{k_{\mathrm{L}}^2}\sum_{j=1}^{k_{\mathrm{L}}-1}\sum_{i=j+1}^{k_{\mathrm{L}}}\mathbb{E}\{\eta(X_{(i)}) - \eta(x)\}\mathbb{E}\{\eta(X_{(j)}) - \eta(x)\} \\
&\quad= \frac{n^2}{k_{\mathrm{L}}^2}\sum_{j=1}^{k_{\mathrm{L}}-1}\sum_{i=j+1}^{k_{\mathrm{L}}}\int_{u_2\in\mathbb{R}^d}\int_{u_1\in\mathbb{R}^d}\gamma(x,u_1,u_2)p_{\|u_2\|}^{n-1}(j-1)p_{\|u_1\|}^{n-1}(i-1)\,du_1\,du_2 \\
&\quad= \frac{n^2}{k_{\mathrm{L}}^2}\sum_{j=1}^{k_{\mathrm{L}}-1}\int_{u_2\in\mathbb{R}^d}\int_{u_1\in\mathbb{R}^d}\gamma(x,u_1,u_2)p_{\|u_2\|}^{n-1}(j-1)\{q_{\|u_1\|}^{n-1}(k_{\mathrm{L}}) - q_{\|u_1\|}^{n-1}(j)\}\,du_1\,du_2.
\end{aligned}$$

Observe that

$$\begin{aligned}
&\sum_{j=2}^{k_{\mathrm{L}}}p_{\|u_2\|}^{n-1}(j-1)q_{\|u_1\|}^{n-1}(j-1) + \sum_{j=1}^{k_{\mathrm{L}}-1}p_{\|u_2\|}^{n-1}(j-1)\{q_{\|u_1\|}^{n-1}(k_{\mathrm{L}}) - q_{\|u_1\|}^{n-1}(j)\} \\
&\quad= \sum_{j=1}^{k_{\mathrm{L}}-1}p_{\|u_2\|}^{n-1}(j-1)q_{\|u_1\|}^{n-1}(k_{\mathrm{L}})\{1+o(1)\} = q_{\|u_2\|}^{n-1}(k_{\mathrm{L}}-1)q_{\|u_1\|}^{n-1}(k_{\mathrm{L}})\{1+o(1)\}
\end{aligned}$$

as $n\to\infty$, uniformly for $x\in\mathcal{R}_n$ and $k_{\mathrm{L}}\in K_\beta$. Thus

$$\begin{aligned}
&\frac{2}{k_{\mathrm{L}}^2}\sum_{j=2}^{k_{\mathrm{L}}}\sum_{i=1}^{j-1}\mathbb{E}\{\eta(X_{(i)}) - \eta(x)\}\mathbb{E}\{\eta(X_{(j)}) - \eta(x)\} \\
&\quad= \frac{n^2}{k_{\mathrm{L}}^2}\int_{u_2\in\mathbb{R}^d}\int_{u_1\in\mathbb{R}^d}\gamma(x,u_1,u_2)q_{\|u_2\|}^{n-1}(k_{\mathrm{L}}-1)q_{\|u_1\|}^{n-1}(k_{\mathrm{L}})\,du_1\,du_2\{1+o(1)\} \\
&\quad= \Big\{\frac{n}{k_{\mathrm{L}}}\int_{\|u_1\|\le 1/b_n}\{\eta(x+u_1)-\eta(x)\}\bar{f}(x+u_2)\,du_1\Big\}^2\{1+o(1)\} \\
&\quad= \Big(\frac{k_{\mathrm{L}}}{n\bar{f}(x)}\Big)^{4/d}a(x)^2\{1+o(1)\} \qquad (3.34)
\end{aligned}$$

as $n\to\infty$, uniformly for $x\in\mathcal{R}_n$ and $k_{\mathrm{L}}\in K_\beta$, where $b_n = \big(\frac{(n-1)\bar{f}(x)a_d}{k_{\mathrm{L}}}\big)^{1/d}$. Using (3.33)

and (3.34) we have that

$$\frac{2}{k_{\mathrm{L}}^2}\sum_{j=2}^{k_{\mathrm{L}}}\sum_{i=1}^{j-1}\mathbb{E}\big[\{\eta(X_{(i)})\}\{\eta(X_{(j)})\}\big]-\mathbb{E}\big[\{\eta(X_{(i)})\}\mathbb{E}\{\eta(X_{(j)})\}\big]$$
$$=\frac{2C_d}{c_d}\Big(\frac{k_{\mathrm{L}}}{n\bar{f}(x)}\Big)^{4/d}a(x)^2\{1+o(1)\}-\Big(\frac{k_{\mathrm{L}}}{n\bar{f}(x)}\Big)^{4/d}a(x)^2\{1+o(1)\},\qquad(3.35)$$

uniformly for $x\in\mathcal{R}_n$ and $k_{\mathrm{L}}\in K_\beta$. As a final observation, note that

$$\frac{2C_d}{c_d}=\frac{2\int_{\|v\|\le 1}\|v\|^{d+2}v_1^2\,dv}{\int_{\|v\|\le 1}v_1^2\,dv}=1.$$

Therefore, the two dominant terms in (3.35) cancel and we deduce that (3.30) holds. This concludes Step 3.

Step 4: Recall that $\epsilon_n:=\frac{1}{\log^{1/2}(n^\beta)}$. We show that

$$\sup_{k_{\mathrm{L}}\in K_\beta}\sup_{x\in\mathcal{R}_n\setminus\mathcal{S}_n^{\epsilon_n}}|\mathbb{P}\{\hat{S}_n(x)<1/2\}-\mathbb{1}_{\{\eta(x)<1/2\}}|=O(n^{-M}),$$

for each $M>0$, as $n\to\infty$.

First note that, since $\|\dot{\eta}(x)\|$ is bounded away from zero on the whole of $\mathcal{S}$, there exists $c_1>0$ and $\epsilon_0>0$ (independent of n) such that, for all $\epsilon<\epsilon_0$,

$$\inf_{x\in\mathcal{R}_n\setminus\mathcal{S}_n^{\epsilon}}|\eta(x)-1/2|\ge c_1\epsilon.$$

Moreover, by Bernstein's inequality,

$$\sup_{k_{\mathrm{L}}\in K_\beta}\sup_{x\in\mathcal{R}_n}q^n_{\epsilon_n/2}(k_{\mathrm{L}})\le\sup_{k_{\mathrm{L}}\in K_\beta}\sup_{x\in\mathcal{R}_n}\exp\Big(-\frac{(np_{\epsilon_n/2}-k_{\mathrm{L}})^2}{2np_{\epsilon_n/2}(1-p_{\epsilon_n/2})+2(np_{\epsilon_n/2}-k_{\mathrm{L}})/3}\Big)$$
$$\le\sup_{k_{\mathrm{L}}\in K_\beta}\sup_{x\in\mathcal{R}_n}\exp\Big(-\frac{np_{\epsilon_n/2}}{3}\Big(1-\frac{k_{\mathrm{L}}}{np_{\epsilon_n/2}}\Big)^2\Big)=O(n^{-M}),$$

for each $M>0$, as $n\to\infty$, since, by assumption **(B.3)** there exists $\kappa>0$ such that

$$\inf_{x\in\mathcal{R}_n}np_{\epsilon_n/2}\ge\kappa k_{\mathrm{L}}\log^{3d/2}(n^\beta),$$

for all n sufficiently large. It follows that

$$\inf_{\substack{x\in\mathcal{R}_n\setminus\mathcal{S}_n^{\epsilon_n}:\\ \eta(x)\ge 1/2}}\mu_n(x)-1/2\ge\inf_{\substack{x\in\mathcal{R}_n\setminus\mathcal{S}_n^{\epsilon_n}:\\ \eta(x)\ge 1/2}}\frac{1}{k_{\mathrm{L}}}\sum_{i=1}^{k_{\mathrm{L}}}\mathbb{P}\{Y_{(i)}=1\cap\|X_{(i)}-x\|\le\epsilon_n/2\}-1/2$$
$$\ge\Big\{\frac{1}{2}+c_1\epsilon_n/2\Big\}\{1-q^n_{\epsilon_n/2}(k_{\mathrm{L}})\}-1/2\ge\frac{c_1\epsilon_n}{4}.$$

Similarly

$$\sup_{\substack{x\in\mathcal{R}_n\setminus\mathcal{S}_n^{\epsilon_n}:\\ \eta(x)\le 1/2}} \mu_n(x) - 1/2 \le -\frac{c_1\epsilon_n}{4}.$$

Now, conditioned on X^n, $\hat{S}_n(x)$ is the sum of $k_{\rm L}(x)$ independent terms. Therefore, by Hoeffding's inequality,

$$\begin{aligned}
&\sup_{x\in\mathcal{R}_n\setminus\mathcal{S}_n^{\epsilon_n}} |\mathbb{P}\{\hat{S}_n(x) < 1/2\} - \mathbb{1}_{\{\eta(x)\le 1/2\}}| \\
&\qquad = \sup_{x\in\mathcal{R}_n\setminus\mathcal{S}_n^{\epsilon_n}} |\mathbb{E}\{\mathbb{P}\{\hat{S}_n(x) < 1/2|X^n\} - \mathbb{1}_{\{\eta(x)\le 1/2\}}| \\
&\qquad \le \sup_{x\in\mathcal{R}_n\setminus\mathcal{S}_n^{\epsilon_n}} \mathbb{E}\exp(-2k_{\rm L}\{\hat{\mu}_n(x, X^n) - 1/2\}^2) \\
&\qquad = \sup_{x\in\mathcal{R}_n\setminus\mathcal{S}_n^{\epsilon_n}} \exp(-2k_{\rm L}\{\mu_n(x) - 1/2\}^2) + R_5 \\
&\qquad \le \sup_{x\in\mathcal{R}_n\setminus\mathcal{S}_n^{\epsilon_n}} \exp\Big(-\frac{1}{8}c_1^2 k_{\rm L}(x)\epsilon_n^2\Big) + R_5 = O(n^{-M}),
\end{aligned}$$

where we show in Step 7 that $|R_5| = O(n^{-M})$, for each $M > 0$. We conclude that

$$\sup_{k_{\rm L}\in K_\beta} \int_{\mathcal{R}_n\setminus\mathcal{S}_n^{\epsilon_n}} [\mathbb{P}\{\hat{S}_n(x) < 1/2\} - \mathbb{1}_{\{\eta(x)\le 1/2\}}]\{\eta(x) - 1/2\}\bar{f}(x)\,dx = O(n^{-M}),$$

for each $M > 0$, as $n \to \infty$. This completes Step 4.

Step 5: Recall that $x_0^t = x_0 + t\dot{\eta}(x_0)/\|\dot{\eta}(x_0)\|$, and let

$$\psi(x) := \{2\eta(x) - 1\}\bar{f}(x) = \pi_1 f_1(x) - \pi_2 f_2(x).$$

We show that

$$\begin{aligned}
&\int_{\mathcal{S}_n^{\epsilon_n}\cap\mathcal{R}_n} [\mathbb{P}\{\hat{S}_n(x) < 1/2\} - \mathbb{1}_{\{\eta(x)<1/2\}}]\{2\eta(x) - 1\}\bar{f}(x)\,dx \\
&\qquad = \int_{\mathcal{S}\cap\mathcal{R}_n}\int_{-\epsilon_n}^{\epsilon_n} \psi(x_0^t)[\mathbb{P}\{\hat{S}_n(x_0^t) < 1/2\} - \mathbb{1}_{\{t<0\}}]\,dt\,d\mathrm{Vol}^{d-1}(x_0)\{1 + o(1)\} \\
&\qquad\qquad + o\big(\sup_{x\in\mathcal{R}_n^c}\{\bar{f}(x)\}\big),
\end{aligned}$$

uniformly for $k_{\rm L} \in K_\beta$.

Recall further $\mathcal{S}_n^{\epsilon\epsilon} = \{x \in \mathbb{R}^d : \eta(x) = 1/2 \text{ and } \mathrm{dist}(x, \mathcal{S}_n) < \epsilon\}$. Now, the map

$$\phi(x_0, t\dot{\eta}(x_0)/\|\dot{\eta}(x_0)\|) = x_0^t$$

is a diffeomorphism from $\{(x_0, t\dot{\eta}(x_0)/\|\dot{\eta}(x_0)\|) : x_0 \in \mathcal{S}_n^{\epsilon_n\epsilon_n}, |t| < \epsilon_n\}$ onto $\mathcal{S}_n^{\epsilon_n}$. Furthermore, for large n, and $|t| \le \epsilon_n$, $\mathrm{sgn}\{\eta(x_0^t) - 1/2\} = \mathrm{sgn}(t)$. The pullback of the

d-form dx is given at $(x_0, t\dot{\eta}(x_0)/\|\dot{\eta}(x_0)\|)$ by

$$\det\dot{\phi}(x_0, t\dot{\eta}(x_0)/\|\dot{\eta}(x_0)\|)\, dt\, d\mathrm{Vol}^{d-1}(x_0) = \{1 + o(1)\}\, dt\, d\mathrm{Vol}^{d-1}(x_0),$$

where the error is uniform for $x_0 \in \mathcal{S}_n^{\epsilon_n \epsilon_n}$, $|t| \leq \epsilon_n$. Therefore, since, for each n, $\mathcal{S}_n^{\epsilon_n}$ is compact, by Weyl's tube formula (Gray, 2004), we have that

$$\begin{aligned}\int_{\mathcal{S}_n^{\epsilon_n} \cap \mathcal{R}_n} & [\mathbb{P}\{\hat{S}_n(x) < 1/2\} - \mathbb{1}_{\{\eta(x)<1/2\}}]\{\eta(x) - 1/2\}\bar{f}(x)\, dx \\ & = \int_{\mathcal{S}^{\epsilon_n \epsilon_n}} \int_{-\epsilon_n}^{\epsilon_n} \psi(x_0^t)[\mathbb{P}\{\hat{S}_n(x_0^t) < 1/2\} - \mathbb{1}_{\{t<0\}}]\, dt\, d\mathrm{Vol}^{d-1}(x_0)\{1 + o(1)\}.\end{aligned}$$

Moreover

$$\begin{aligned}\Big| \int_{\mathcal{S}_n^{\epsilon_n}} & [\mathbb{P}\{\hat{S}_n(x) < 1/2\} - \mathbb{1}_{\{\eta(x)<1/2\}}]\{\eta(x) - 1/2\}\bar{f}(x)\, dx \\ & \qquad - \int_{\mathcal{S}_n^{\epsilon_n} \cap \mathcal{R}_n} [\mathbb{P}\{\hat{S}_n(x) < 1/2\} - \mathbb{1}_{\{\eta(x)<1/2\}}]\{\eta(x) - 1/2\}\bar{f}(x)\, dx \Big| \\ & \leq \sup_{x \in \mathcal{R}_n^c}\{\bar{f}(x)\} \int_{\mathcal{S}_n^{\epsilon_n} \cap \mathcal{R}_n^c} dx = O\big(\sup_{x \in \mathcal{R}_n^c}\{\bar{f}(x)\}\epsilon_n^2\big).\end{aligned}$$

Then, by another application of Weyl's tube formula, we have that

$$\begin{aligned}\Big| \int_{\mathcal{S}_n^{\epsilon_n \epsilon_n}} & \int_{-\epsilon_n}^{\epsilon_n} \psi(x_0^t)[\mathbb{P}\{\hat{S}_n(x_0^t) < 1/2\} - \mathbb{1}_{\{t<0\}}]\, dt\, d\mathrm{Vol}^{d-1}(x_0) \\ & \qquad - \int_{\mathcal{S} \cap \mathcal{R}_n} \int_{-\epsilon_n}^{\epsilon_n} \psi(x_0^t)[\mathbb{P}\{\hat{S}_n(x_0^t) < 1/2\} - \mathbb{1}_{\{t<0\}}]\, dt\, d\mathrm{Vol}^{d-1}(x_0) \Big| \\ & \leq \sup_{x \in \mathcal{R}_n^c}\{\bar{f}(x)\} \int_{\mathcal{S}_n^{\epsilon_n \epsilon_n} \cap \mathcal{R}_n^c} \int_{-\epsilon_n}^{\epsilon_n} dt\, d\mathrm{Vol}^{d-1}(x_0) = O\big(\sup_{x \in \mathcal{R}_n^c}\{\bar{f}(x)\}\epsilon_n^2\big).\end{aligned}$$

This completes step 5.

Step 6: The last step in the main argument is to show that

$$\begin{aligned}\int_{\mathcal{S} \cap \mathcal{R}_n} & \int_{-\epsilon_n}^{\epsilon_n} \psi(x_0^t)[\mathbb{P}\{\hat{S}_n(x_0^t) < 1/2\} - \mathbb{1}_{\{t<0\}}]\, dt\, d\mathrm{Vol}^{d-1}(x_0) \\ & = \Big\{ \int_{S \cap \mathcal{R}_n} \frac{\bar{f}(x_0)}{4k_{\mathrm{L}}(x_0)\|\dot{\eta}(x_0)\|}\, d\mathrm{Vol}^{d-1}(x_0) \\ & \qquad + \frac{1}{n^{4/d}} \int_{S \cap \mathcal{R}_n} \frac{\bar{f}(x_0)^{1-4/d} k_{\mathrm{L}}(x_0)^{4/d}}{\|\dot{\eta}(x_0)\|} a(x_0)^2\, d\mathrm{Vol}^{d-1}(x_0) \Big\}\{1 + o(1)\}\end{aligned}$$

as $n \to \infty$, uniformly for $k_{\mathrm{L}} \in K_\beta$.

First observe that

$$\int_{\mathcal{R}_n\cap\mathcal{S}}\int_{-\epsilon_n}^{\epsilon_n}\psi(x_0^t)[\mathbb{P}\{\hat{S}_n(x_0^t)<1/2\}-\mathbb{1}_{\{t<0\}}]\,dt\,d\mathrm{Vol}^{d-1}(x_0)$$
$$=\int_{\mathcal{R}_n\cap\mathcal{S}}\int_{-\epsilon_n}^{\epsilon_n}t\|\dot{\psi}(x_0)\|[\mathbb{P}\{\hat{S}_n(x_0^t)<1/2\}-\mathbb{1}_{\{t<0\}}]\,dt\,d\mathrm{Vol}^{d-1}(x_0)\{1+o(1)\}.$$

Now, write $\mathbb{P}\{\hat{S}_n(x_0^t)<1/2\}-\mathbb{1}_{\{t<0\}}=\mathbb{E}[\mathbb{P}\{\hat{S}_n(x_0^t)<1/2|X^n\}-\mathbb{1}_{\{t<0\}}]$. Note that, given X^n, $\hat{S}_n(x)=\frac{1}{k_{\mathrm{L}}(x)}\sum_{i=1}^{k_{\mathrm{L}}(x)}\mathbb{1}_{\{Y_{(i)}=1\}}$ is the sum of $k_{\mathrm{L}}(x)$ independent Bernoulli variables, satisfying $\mathbb{P}(Y_{(i)}=1|X^n)=\eta(X_{(i)})$. Therefore, by the Berry–Esseen Theorem, there exists $C_1>0$, such that

$$\begin{aligned}\sup_{z\in\mathbb{R}}\Big|\mathbb{P}\Big\{\frac{\hat{S}_n(x_0^t)-\hat{\mu}_n(x_0^t,x^n)}{\hat{\sigma}_n(x_0^t,x^n)}<z\Big|X^n=x^n\Big\}-\Phi(z)\Big|&\\ \le\frac{C_1\sum_{i=1}^{k_{\mathrm{L}}(x_0^t)}\eta(x_{(i)})\{1-\eta(x_{(i)})\}\{2\eta(x_{(i)})^2-2\eta(x_{(i)})+1\}}{k_{\mathrm{L}}(x_0^t)^3\hat{\sigma}_n^3(x_0^t,x^n)}&\\ \le\frac{C_1}{4k_{\mathrm{L}}(x_0^t)^2\sigma_n^3(x_0^t,x^n)},&\end{aligned}$$

where Φ denotes the standard normal distribution function. Thus

$$\Big|\mathbb{P}\{\hat{S}_n(x_0^t)<1/2|X^n=x^n\}-\Phi\Big(\frac{1/2-\hat{\mu}_n(x_0^t,x^n)}{\hat{\sigma}_n(x_0^t,x^n)}\Big)\Big|\le\frac{C_1}{4k_{\mathrm{L}}(x_0^t)^2\sigma_n^3(x_0^t,x^n)}.\quad(3.36)$$

It follows that

$$\begin{aligned}&\int_{-\epsilon_n}^{\epsilon_n}t\|\dot{\psi}(x_0)\|[\mathbb{P}\{\hat{S}_n(x_0^t)<1/2\}-\mathbb{1}_{\{t<0\}}]\,dt\\ &\quad=\int_{-\epsilon_n}^{\epsilon_n}t\|\dot{\psi}(x_0)\|\Big\{\Phi\big(2k_{\mathrm{L}}(x_0)^{1/2}\{1/2-\mu_n(x_0^t)\}\big)-\mathbb{1}_{\{t<0\}}\Big\}\,dt+R_6(x_0)\\ &\quad=\int_{-\epsilon_n}^{\epsilon_n}t\|\dot{\psi}(x_0)\|\Big\{\Phi\Big(-2k_{\mathrm{L}}(x_0)^{1/2}\Big\{t\|\dot{\eta}(x_0)\|+\Big(\frac{k_{\mathrm{L}}(x_0)}{n\bar{f}(x_0)}\Big)^{2/d}a(x_0)\Big\}\Big)-\mathbb{1}_{\{t<0\}}\Big\}\,dt\\ &\qquad+R_6(x_0)+R_7(x_0),\end{aligned}$$

where we have used the fact that, for all $k_{\mathrm{L}}\in K_\beta$, and all n sufficiently large, $\sup_{|t|\le\epsilon_n}|k_{\mathrm{L}}(x_0^t)-k_{\mathrm{L}}(x_0)|\le 1$. We show in Step 7 that

$$\Big|\int_{\mathcal{R}_n\cap\mathcal{S}}R_6(x_0)+R_7(x_0)\,d\mathrm{Vol}^{d-1}(x_0)\Big|=o(\gamma_n(k_{\mathrm{L}})).$$

Then, substituting $r = 2k_{\mathrm{L}}(x_0)^{1/2}t$, we see that

$$\begin{aligned}
&\int_{-\epsilon_n}^{\epsilon_n} t\|\dot{\psi}(x_0)\|\Big[\Phi\Big(-2k_{\mathrm{L}}(x_0)^{1/2}\Big\{t\|\dot{\eta}(x_0)\| + \Big(\frac{k_{\mathrm{L}}(x_0)}{n\bar{f}(x_0)}\Big)^{2/d}a(x_0)\Big\}\Big) - \mathbb{1}_{\{t<0\}}\Big]\,dt \\
&= \frac{1}{4k_{\mathrm{L}}(x_0)}\int_{-2k_{\mathrm{L}}(x_0)^{1/2}\epsilon_n}^{2k_{\mathrm{L}}(x_0)^{1/2}\epsilon_n} r\|\dot{\psi}(x_0)\|\Big\{\Phi\Big(-r\|\dot{\eta}(x_0)\| - 2k_{\mathrm{L}}(x_0)^{1/2}\Big(\frac{k_{\mathrm{L}}(x_0)}{n\bar{f}(x_0)}\Big)^{2/d}a(x_0)\Big) \\
&\qquad\qquad - \mathbb{1}_{\{r<0\}}\Big\}\,dr \\
&= \Big\{\frac{\bar{f}(x_0)}{4k_{\mathrm{L}}(x_0)\|\dot{\eta}(x_0)\|} + \Big(\frac{k_{\mathrm{L}}(x_0)}{n\bar{f}(x_0)}\Big)^{4/d}\frac{\bar{f}(x_0)a(x_0)^2}{\|\dot{\eta}(x_0)\|}\Big\}\{1+o(1)\}.
\end{aligned}$$

The conclusion follows by integrating x_0 over $\mathcal{R}_n \cap \mathcal{S}$.

Step 7: It remains to bound the error terms R_1, R_2, R_5, R_6, and R_7.

To bound R_1: write

$$R_1 = \frac{1}{k_{\mathrm{L}}}\sum_{i=1}^{k_{\mathrm{L}}}\mathbb{E}\eta(X_{(i)}) - \eta(x) - \mathbb{E}\{(X_{(i)}-x)^T\dot{\eta}(x)\} - \frac{1}{2}\mathbb{E}\{(X_{(i)}-x)^T\ddot{\eta}(x)(X_{(i)}-x)\}.$$

By a Taylor expansion, for all $\epsilon > 0$, there exists $\delta = \delta_\epsilon > 0$, such that

$$|\eta(z) - \eta(x) - (z-x)^T\dot{\eta}(x) - \frac{1}{2}(z-x)^T\ddot{\eta}(x)(z-x)| \le \epsilon\|z-x\|^2,$$

for all $\|z-x\| < \delta$. Hence

$$\begin{aligned}
|R_1| \le \epsilon\frac{1}{k_{\mathrm{L}}}\sum_{i=1}^{k_{\mathrm{L}}}\mathbb{E}\{\|X_{(i)}-x\|^2\mathbb{1}_{\{\|X_{(k_{\mathrm{L}})}-x\|\le\delta\}}\} + 2\mathbb{P}\{\|X_{(k_{\mathrm{L}})}-x\| > \delta\} \\
+ (1+D_1)\mathbb{E}\{\|X_{(k_{\mathrm{L}})}-x\|\mathbb{1}_{\{\|X_{(k_{\mathrm{L}})}-x\|>\delta\}}\} \\
+ (1+D_2)\mathbb{E}\{\|X_{(k_{\mathrm{L}})}-x\|^2\mathbb{1}_{\{\|X_{(k_{\mathrm{L}})}-x\|>\delta\}}\},
\end{aligned}$$

where $D_1 := \sup_{x_0\in\mathcal{S}}\|\dot{\eta}(x_0)\|$, and $D_2 := \sup_{x_0\in\mathcal{S}}\lambda_{\max}\{\ddot{\eta}(x_0)\}$ (here $\lambda_{\max}(\cdot)$ denotes the largest eigenvalue of a matrix). Now, by similar arguments to those leading to (3.28), we have that

$$\frac{\epsilon}{k_{\mathrm{L}}}\sum_{i=1}^{k_{\mathrm{L}}}\mathbb{E}(\|X_{(i)}-x\|^2\mathbb{1}_{\{\|X_{(k_{\mathrm{L}})}-x\|\le\delta\}}) = \epsilon\Big(\frac{k_{\mathrm{L}}}{na_d\bar{f}(x)}\Big)^{2/d}\frac{d}{d+2}\{1+o(1)\}.$$

Moreover,

$$\mathbb{P}\{\|X_{(k_{\mathrm{L}})}-x\| > \delta\} = q_\delta^n(k_{\mathrm{L}}) = O(n^{-M}),$$

by (3.27) in Step 1. For the remaining terms, note that

$$\mathbb{E}\{\|X_{(k_{\mathrm{L}})}-x\|^2\mathbb{1}_{\{\|X_{(k_{\mathrm{L}})}-x\|>\delta\}}\} = \mathbb{P}\{\|X_{(k_{\mathrm{L}})}-x\|>\delta\} + \int_{\delta^2}^{\infty} \mathbb{P}\{\|X_{(k_{\mathrm{L}})}-x\|>\sqrt{t}\}dt$$
$$= q_\delta^n(k_{\mathrm{L}}) + \int_{\delta^2}^{\infty} q_{\sqrt{t}}^n(k_{\mathrm{L}})\,dt.$$

Now, for all $t_0 > \delta^2$, there exists $c_3 > 0$, such that $np_\delta - k_{\mathrm{L}} \geq c_3\delta^d n$ for all $x \in \mathcal{S}\cap\mathcal{R}_n$, thus

$$\sup_{x\in\mathcal{S}\cap\mathcal{R}_n}\sup_{k_{\mathrm{L}}\in K_\beta}\Big\{q_\delta^n(k_{\mathrm{L}}) + \int_{\delta^2}^{t_0} q_{\sqrt{t}}^n(k_{\mathrm{L}})\,dt\Big\} \leq (1+t_0)\exp(-2c_2^2 n\delta^{2d}) = O(n^{-M}).$$

Furthermore, using Assumption **(B.4)(ρ)**, by Bennett's inequality, for all n sufficiently large and all t_0, there exist c_4, $c_5 > 0$ such that, for all $t > t_0$,

$$\sup_{k_{\mathrm{L}}\in K_\beta}\sup_{x\in\mathcal{S}\cap\mathcal{R}_n} q_{\sqrt{t}}^n(k_{\mathrm{L}}) \leq (1+c_4 t^{\rho/2})^{-c_5 n}.$$

Then, using Markov's inequality to deal with the final term, we conclude that $|R_1| = o\big(\big(\frac{k_{\mathrm{L}}}{nf(x)}\big)^{2/d}\{1+|a(x)|\}\big)$. Note further that, with only simple modifications, we have also shown (3.23), which bounds R_2.

To bound R_5: Observe that

$$|R_5| \leq \sup_{x\in\mathcal{R}_n\setminus\mathcal{S}^{\epsilon_n}} |\mathbb{E}\exp(-2k_{\mathrm{L}}\{\hat{\mu}_n(x,X^n)-1/2\}^2) - \exp(-2k_{\mathrm{L}}\{\mu_n(x)-1/2\}^2)|.$$

Let $\hat{\theta}(x) := -(2k_{\mathrm{L}})^{1/2}\{\hat{\mu}_n(x,X^n)-1/2\}$ and $\theta(x) := \mathbb{E}\hat{\theta}(x) = -(2k_{\mathrm{L}})^{1/2}\mathbb{E}\{\hat{\mu}_n(x,X^n)-1/2\}$. By Step 3, given $\epsilon > 0$ sufficiently small, for all n sufficiently large we have that, for all $k_{\mathrm{L}} \in K_\beta$, $x \in \mathcal{R}_n \setminus \mathcal{S}^{\epsilon_n}$,

$$\mathbb{E}\hat{\theta}(x)^2 - \theta(x)^2 = 2k_{\mathrm{L}}\mathrm{Var}\{\hat{\mu}(x,X^n)\} \leq \epsilon\Big\{1 + k_{\mathrm{L}}(x)a(x)^2\Big(\frac{k_{\mathrm{L}}(x)}{n\bar{f}(x)}\Big)^{4/d}\Big\}.$$

By a Taylor expansion, it follows that

$$|\mathbb{E}e^{-\hat{\theta}(x)^2} - e^{-\theta(x)^2}| \leq \mathbb{E}|e^{-\hat{\theta}(x)^2} - e^{-\theta(x)^2}|$$
$$= \mathbb{E}|\hat{\theta}(x)^2 - \theta(x)^2|e^{-\theta(x)^2}\{1+o(1)\} = O(n^{-M}),$$

for each $M > 0$.

To bound R_6: using the θ notation introduced above, we decompose R_6 as follows:

$$R_6 := \int_{\mathcal{R}_n\cap\mathcal{S}} R_6(x_0)\,d\mathrm{Vol}^{d-1}(x_0) = R_{61} + R_{62},$$

where

$$R_{61} := \int_{\mathcal{R}_n \cap \mathcal{S}} \int_{-\epsilon_n}^{\epsilon_n} t \|\dot{\psi}(x_0)\| \Big[\mathbb{P}\{\hat{S}_n(x_0^t) < 1/2\} - \mathbb{E}\Phi\big(\hat{\theta}(x_0^t)\big) \Big] \, dt \, d\mathrm{Vol}^{d-1}(x_0),$$

and

$$R_{62} := \int_{\mathcal{R}_n \cap \mathcal{S}} \int_{-\epsilon_n}^{\epsilon_n} t \|\dot{\psi}(x_0)\| \Big\{ \mathbb{E}\Phi\big(\hat{\theta}(x_0^t)\big) - \Phi\big(\theta(x_0^t)\big) \Big\} \, dt \, d\mathrm{Vol}^{d-1}(x_0).$$

To bound R_{61}: Write

$$R_{61} := \int_{\mathcal{R}_n \cap \mathcal{S}} \int_{-\epsilon_n}^{\epsilon_n} t \|\dot{\psi}(x_0)\| \mathbb{E}\big[\mathbb{P}\{\hat{S}_n(x_0^t) < 1/2 | X^n\} - \Phi\big(\hat{\theta}(x_0^t)\big) \big] \, dt \, d\mathrm{Vol}^{d-1}(x_0).$$

By Step 1 and 2, there exists c_1 and C_2, such that, for all $x_0 \in \mathcal{S} \cap \mathcal{R}_n$ and $|t| \in (C_2(k_{\mathrm{L}}(x_0^t)/n)^{4/d}, \epsilon_n)$,

$$\hat{\theta}(x_0^t) \geq c_1 k_{\mathrm{L}}(x_0^t)^{1/2} |t|.$$

Hence, by the non-uniform version of the Berry–Esseen Theorem (Petrov, 1975), we have that

$$\mathbb{E}\big[\mathbb{P}\{\hat{S}_n(x_0^t) < 1/2 | X^n\} - \Phi\big(\hat{\theta}(x_0^t)\big) \big] \leq \frac{C_1}{k_{\mathrm{L}}(x_0^t)^{1/2}} \frac{1}{1 + c_1^3 k_{\mathrm{L}}(x_0^t)^{3/2} |t|^3},$$

for all $|t| \in (C_2(k_{\mathrm{L}}(x_0^t)/n)^{2/d}, \epsilon_n)$. Now, by the Berry–Esseen bound in (3.36) and Step 2, we have, for all $|t| \in (0, C_2(k_{\mathrm{L}}/n)^{2/d})$,

$$\mathbb{E}\big[\mathbb{P}\{\hat{S}_n(x_0^t) < 1/2 | X^n\} - \Phi\big(\hat{\theta}(x_0^t)\big) \big] \leq \frac{C_1'}{k_{\mathrm{L}}(x_0^t)^{1/2}}.$$

It follows that

$$\begin{aligned}
|R_{61}| &\leq C_1' \int_{\mathcal{R}_n \cap \mathcal{S}} \int_{-C_2(k_{\mathrm{L}}(x_0^t)/n)^{2/d}}^{C_2(k_{\mathrm{L}}(x_0^t)/n)^{2/d}} \frac{t \|\dot{\psi}(x_0)\|}{k_{\mathrm{L}}(x_0^t)^{1/2}} \, dt \, d\mathrm{Vol}^{d-1}(x_0) \\
&\quad + C_1' \int_{\mathcal{R}_n \cap \mathcal{S}} \int_{|t| \in (C_2(k_{\mathrm{L}}(x_0^t)/n)^{2/d}, \epsilon_n)} \frac{|t| \|\dot{\psi}(x_0)\|}{1 + c_1^3 k_{\mathrm{L}}(x_0^t)^2 |t|^3} \, dt \, d\mathrm{Vol}^{d-1}(x_0) \\
&= C_1' \int_{\mathcal{R}_n \cap \mathcal{S}} \int_{-C_2(k_{\mathrm{L}}(x_0)/n)^{2/d}}^{C_2(k_{\mathrm{L}}(x_0)/n)^{2/d}} \frac{t \|\dot{\psi}(x_0)\|}{k_{\mathrm{L}}(x_0)^{1/2}} \, dt \, d\mathrm{Vol}^{d-1}(x_0) \\
&\quad + C_1' \int_{\mathcal{R}_n \cap \mathcal{S}} \int_{|t| \in (C_2(k_{\mathrm{L}}(x_0)/n)^{2/d}, \epsilon_n)} \frac{|t| \|\dot{\psi}(x_0)\|}{1 + c_1^3 k_{\mathrm{L}}(x_0)^2 |t|^3} \, dt \, d\mathrm{Vol}^{d-1}(x_0) \{1 + o(1)\} \\
&= o(\gamma_n(k_{\mathrm{L}})),
\end{aligned}$$

uniformly for $k_{\mathrm{L}} \in K_\beta$.

To bound R_{62}: By Step 3, given $\epsilon > 0$ sufficiently small, for all sufficiently large n,

we have that, for all $k_{\mathrm{L}} \in K_\beta$, $x_0 \in \mathcal{S} \cap \mathcal{R}_n$ and all $|t| \in [\epsilon(k_{\mathrm{L}}(x_0^t)/n)^{2/d}, \epsilon_n]$,

$$\mathbb{E}\big|\hat{\theta}(x_0^t) - \theta(x_0^t)\big| \leq 2k_{\mathrm{L}}(x_0^t)^{1/2}\mathrm{Var}\{\hat{\mu}(x_0^t, X^n)\}^{1/2} \leq \epsilon^2\Big\{1 + k_{\mathrm{L}}^{1/2}a(x_0^t)\Big(\frac{k_{\mathrm{L}}(x_0^t)}{n\bar{f}(x_0^t)}\Big)^{2/d}\Big\}$$
$$\leq \epsilon^2 + \frac{\epsilon a(x_0^t)k_{\mathrm{L}}(x_0^t)^{1/2}|t|}{\bar{f}(x_0^t)^{2/d}}.$$

It follows that, for large n, $|t| \in [\epsilon(k_{\mathrm{L}}(x_0)/n)^{2/d}, \epsilon_n]$ and all $k_{\mathrm{L}} \in K_\beta$,

$$\mathbb{E}|\Phi(\hat{\theta}) - \Phi(\theta)| \leq \Big\{\epsilon^2 + \frac{\epsilon a(x_0)k_{\mathrm{L}}(x_0)^{1/2}|t|}{\bar{f}(x_0)^{2/d}}\Big\}$$
$$\phi\Big(-k_{\mathrm{L}}(x_0)^{1/2}|t|\|\dot{\eta}(x_0)\| - k_{\mathrm{L}}(x_0)^{1/2}a(x_0)\Big\{\frac{k_{\mathrm{L}}(x_0)}{n\bar{f}(x_0)}\Big\}^{2/d}\Big),$$

since, by Steps 1 and 2,

$$\theta(x_0^t) = -2k_{\mathrm{L}}^{1/2}\Big[t\|\dot{\eta}(x_0)\| + \Big\{\frac{k_{\mathrm{L}}(x_0)}{n\bar{f}(x_0)}\Big\}^{2/d}a(x_0)\Big] + o\Big(k_{\mathrm{L}}(x_0)^{1/2}a(x_0)\Big\{\frac{k_{\mathrm{L}}(x_0)}{n\bar{f}(x_0)}\Big\}^{2/d}\Big).$$

For $|t| \in [0, \epsilon(k_{\mathrm{L}}(x_0)/n)^{2/d}]$, we use the fact that $\mathbb{E}|\Phi(\hat{\theta}(x_0^t)) - \Phi(\theta(x_0^t))| \leq 1$. Finally, by making the substitution $r = k_{\mathrm{L}}(x_0)^{1/2}t$, we see that

$$R_{62} \leq \int_{\mathcal{S}_n}\int_{|t|\leq\epsilon(k_{\mathrm{L}}(x_0)/n)^{2/d}} |t|\|\dot{\psi}(x_0)\|\,dt\,d\mathrm{Vol}^{d-1}(x_0)$$
$$+ \int_{\mathcal{S}_n}\int_{-\infty}^{\infty} \frac{\|\dot{\psi}(x_0)\|}{k_{\mathrm{L}}(x_0)}\Big\{\epsilon^2 + \frac{\epsilon a(x_0)|r|}{\bar{f}(x_0)^{2/d}}\Big\}$$
$$\phi\Big(-|r|\,\|\dot{\eta}(x_0)\| - \Big\{\frac{k_{\mathrm{L}}(x_0)}{n\bar{f}(x_0)}\Big\}^{2/d}k_{\mathrm{L}}^{1/2}a(x_0)\Big)\,dr\,d\mathrm{Vol}^{d-1}(x_0)$$
$$\leq \epsilon\int_{\mathcal{S}_n}\Big[\Big\{\frac{k_{\mathrm{L}}(x_0)}{n}\Big\}^{4/d}\Big\{1 + \frac{a(x_0)^2}{\bar{f}(x_0)^{4/d}}\Big\} + \frac{1}{k_{\mathrm{L}}(x_0)}\Big]\|\dot{\psi}(x_0)\|\,d\mathrm{Vol}^{d-1}(x_0).$$

for all n sufficiently large, and all $k_{\mathrm{L}} \in K_\beta$.

To bound R_7: write $R_7 := \int_{\mathcal{S}_n} R_7(x_0)\,d\mathrm{Vol}^{d-1}(x_0)$. Let

$$r_x := \frac{-a(x)}{\bar{f}(x)^{2/d}}\frac{k_{\mathrm{L}}(x)^{1/2+2/d}}{n^{2/d}}.$$

By Steps 1 and 2, given $\epsilon > 0$ sufficiently small, for all sufficiently large n, we have that, for all $k_{\mathrm{L}} \in K_\beta$, $x_0 \in \mathcal{S}$ and all $|r| < k_{\mathrm{L}}(x_0)^{1/2}\epsilon_n$,

$$\big|\theta(x_0^{k_{\mathrm{L}}(x_0)^{1/2}r}) + 2\|\dot{\eta}(x_0)\|(r - r_{x_0})\big| \leq \epsilon^2\Big[|r| + k_{\mathrm{L}}(x_0)^{1/2}a(x_0)\Big\{\frac{k_{\mathrm{L}}(x_0)}{n\bar{f}(x_0)}\Big\}^{2/d}\Big].$$

It follows that

$$\Big|\Phi\big(\theta(x_0^t)\big) - \Phi\Big(2k_{\mathrm{L}}(x_0)^{1/2}\Big[-t\|\dot\eta(x_0)\| - a(x_0)\Big\{\frac{k_{\mathrm{L}}(x_0)}{n\bar f(x_0)}\Big\}^{2/d}\Big]\Big)\Big| \\ \leq \epsilon^2\Big[|r| + k_{\mathrm{L}}(x_0)^{1/2}a(x_0)\Big\{\frac{k_{\mathrm{L}}(x_0)}{n\bar f(x_0)}\Big\}^{2/d}\Big]\phi(\|\dot\eta(x_0)\|(r-r_{x_0})),$$

for all $k_{\mathrm{L}} \in K_\beta$, $x_0 \in \mathcal{S}$ and $\epsilon\frac{k_{\mathrm{L}}(x_0)^{1/2+2/d}}{n^{2/d}} \leq |r| < k_{\mathrm{L}}(x_0)^{1/2}\epsilon_n$. Furthermore, for $|r - r_{x_0}| \in [0, \epsilon k_{\mathrm{L}}(x_0)^{1/2+2/d}/n^{2/d}]$ we use the fact that

$$\mathbb{E}|\Phi(\hat\theta(x_0^t)) - \Phi(\theta(x_0^t))| \leq 1.$$

Substituting $r - r_{x_0} = k_{\mathrm{L}}(x_0)^{1/2}t$, it follows that

$$\begin{aligned}|R_7| &\leq \int_{\mathcal{R}_n\cap\mathcal{S}}\int_{|r-r_{x_0}|\leq\epsilon k_{\mathrm{L}}^{1/2+2/d}/n^{2/d}} |r-r_{x_0}|\frac{\|\dot\psi(x_0)\|}{k_{\mathrm{L}}(x_0)}\,dr\,d\mathrm{Vol}^{d-1}(x_0) \\ &\quad + \int_{\mathcal{R}_n\cap\mathcal{S}}\int_{-\infty}^{\infty}\epsilon^2\frac{\|\dot\psi(x_0)\|}{k_{\mathrm{L}}(x_0)}(|r| - r_{x_0})^2\phi(\|\dot\eta(x_0)\|(r-r_{x_0}))\,dr\,d\mathrm{Vol}^{d-1}(x_0) \\ &\leq \epsilon\int_{\mathcal{R}_n\cap\mathcal{S}}\Big[\Big\{\frac{k_{\mathrm{L}}(x_0)}{n}\Big\}^{4/d}\Big\{1 + \frac{a(x_0)^2}{\bar f(x_0)^{4/d}}\Big\} + \frac{1}{k_{\mathrm{L}}(x_0)}\Big]\|\dot\psi(x_0)\|\,d\mathrm{Vol}^{d-1}(x_0).\end{aligned}$$

We conclude that $|R_7| = o(\gamma_n(k_{\mathrm{L}}))$. This completes the proof. □

Proof of Theorem 3.2, part (i). By applying Theorem 3.6 with $k_{\mathrm{L}}(x) = k$ for all $x \in \mathbb{R}^d$, we have that

$$R_{\mathcal{R}_n}(\hat C_n^{k\mathrm{nn}}) - R_{\mathcal{R}_n}(C^{\mathrm{Bayes}}) = \Big\{B_{1,n}\frac{1}{k} + B_{2,n}\Big(\frac{k}{n}\Big)^{4/d}\Big\}\{1 + o(1)\}. \tag{3.37}$$

We show that the contribution from the tail, i.e. for $x \in \mathbb{R}^d \setminus \mathcal{R}_n$, is of smaller order than the terms in (3.37) above. Furthermore, we show that $B_{1,n}$ and $B_{2,n}$ are well approximated by integrals over the whole of $\mathcal{S}$ and that those values are finite.

Let $t_n := \frac{k}{n}\log^{2d}(n/k)$ and fix $\alpha \in (0, 1 - \frac{4(\rho+d)}{d\rho})$. By Markov's inequality and Hölder's inequality, observe that

$$\begin{aligned}&R_{\mathbb{R}^d\setminus\mathcal{R}_n}(\hat C_n^{k\mathrm{nn}}) - R_{\mathbb{R}^d\setminus\mathcal{R}_n}(C^{\mathrm{Bayes}}) \leq \int_{\mathcal{R}_n^c}\bar f(x)\,dx \leq t_n^{\frac{4}{d(1-\alpha)}}\int_{\mathcal{R}_n^c}\bar f(x)^{1-\frac{4}{d(1-\alpha)}}\,dx \\ &\leq t_n^{\frac{4}{d(1-\alpha)}}\Big\{\int_{\mathcal{R}_n^c}(1+\|x\|^\rho)\bar f(x)\,dx\Big\}^{1-\frac{4}{d(1-\alpha)}}\Big\{\int_{\mathcal{R}_n^c}\frac{1}{(1+\|x\|^\rho)^{d(1-\alpha)/4-1}}\,dx\Big\}^{\frac{4}{d(1-\alpha)}} \\ &\qquad = o\Big(\Big(\frac{k}{n}\Big)^{4/d}\Big),\end{aligned} \tag{3.38}$$

uniformly for $k \in K_{\beta,0}$, as $n \to \infty$, since $\rho\{d(1-\alpha)/4 - 1\} > d$.

It remains to show that B_1 and B_2 are finite, and that $B_{1,n}$ and $B_{2,n}$ are well

approximated by the corresponding integrals over the whole region $\mathcal{S}$. Firstly, by Assumption **(B.2)**,

$$B_1 = \int_{\mathcal{S}} \frac{\bar{f}(x_0)}{4\|\dot{\eta}(x_0)\|}\, d\mathrm{Vol}^{d-1}(x_0) \leq \frac{1}{4} \sup_{x_0 \in \mathcal{S}} \Big\{ \frac{1}{\|\dot{\eta}(x_0)\|} \Big\} \int_{\mathcal{S}} \bar{f}(x_0)\, d\mathrm{Vol}^{d-1}(x_0) = O(1).$$

Moreover

$$\begin{aligned} B_1 - B_{1,n} &= \int_{\mathcal{S}\setminus\mathcal{R}_n} \frac{\bar{f}(x_0)}{4\|\dot{\eta}(x_0)\|}\, d\mathrm{Vol}^{d-1}(x_0) \\ &\leq \frac{1}{4} \sup_{x_0 \in \mathcal{S}} \Big\{ \frac{1}{\|\dot{\eta}(x_0)\|} \Big\} \int_{\mathcal{S}\setminus\mathcal{R}_n} \bar{f}(x)\, d\mathrm{Vol}^{d-1}(x_0) = o(1), \end{aligned}$$

To bound B_2: by Assumption **(B.2)** we can define a probability measure on $\mathcal{S}$ by

$$P_{\mathcal{S}}(A) = \frac{\int_A \bar{f}(x)\, d\mathrm{Vol}^{d-1}(x)}{\int_{\mathcal{S}} \bar{f}(x)\, d\mathrm{Vol}^{d-1}(x)},$$

for $A \subseteq \mathcal{S}$. Then, by Jensen's inequality (since $4(\rho+d)/d\rho < 1$), we have that

$$\begin{aligned} B_2 &= \int_{\mathcal{S}} \frac{\bar{f}(x_0)^{1-4/d}}{\|\dot{\eta}(x_0)\|} a(x_0)^2\, d\mathrm{Vol}^{d-1}(x_0) \\ &\leq \sup_{x_0 \in \mathcal{S}} \Big\{ \frac{a(x_0)^2}{\|\dot{\eta}(x_0)\|} \Big\} \int_{\mathcal{S}} \bar{f}(x_0)^{-4/d}\, dP_{\mathcal{S}}(x_0) \int_{\mathcal{S}} \bar{f}(x)\, d\mathrm{Vol}^{d-1}(x) \\ &\leq \sup_{x_0 \in \mathcal{S}} \Big\{ \frac{a(x_0)^2}{\|\dot{\eta}(x_0)\|} \Big\} \Big\{ \int_{\mathcal{S}} \bar{f}(x_0)^{-\rho/(\rho+d)}\, dP_{\mathcal{S}}(x_0) \Big\}^{4(\rho+d)/d\rho} \int_{\mathcal{S}} \bar{f}(x)\, d\mathrm{Vol}^{d-1}(x), \end{aligned}$$

which is finite by assumptions **(B.2)** and **(B.4)(ρ)**. Similarly,

$$\begin{aligned} B_2 - B_{2,n} &= \int_{\mathcal{S}\setminus\mathcal{R}_n} \frac{\bar{f}(x_0)^{1-4/d}}{\|\dot{\eta}(x_0)\|} a(x_0)^2\, d\mathrm{Vol}^{d-1}(x_0) \\ &\leq \sup_{x_0 \in \mathcal{S}} \Big\{ \frac{a(x_0)^2}{\|\dot{\eta}(x_0)\|} \Big\} \int_{\mathcal{S}\setminus\mathcal{R}_n} \bar{f}(x_0)^{1-4/d}\, d\mathrm{Vol}^{d-1}(x_0) \\ &\leq \sup_{x_0 \in \mathcal{S}} \Big\{ \frac{a(x_0)^2}{\|\dot{\eta}(x_0)\|} \Big\} t_n^{\rho/(\rho+d)-4/d} \int_{\mathcal{S}\setminus\mathcal{R}_n} \bar{f}(x_0)^{1-\rho/(\rho+d)}\, d\mathrm{Vol}^{d-1}(x_0) \to 0, \end{aligned}$$

uniformly for $k \in K_{\beta,0}$, as $n \to \infty$. □

Proof of Theorem 3.2, part (ii). Recall that

$$\mathcal{R}_n = \big\{ x \in \mathbb{R}^d : \bar{f}(x) \geq \frac{k}{n} \log^{2d}(n/k) \big\} = \big\{ x \in \mathbb{R}^d : \bar{f}(x) \geq t_n \big\}.$$

In contrast to part (i), the dominant contribution to the excess risk could now arise

from the tail of the distribution. First, by Theorem 3.6, we have that

$$R_{\mathcal{R}_n}(\hat{C}_n^{\mathrm{knn}}) - R_{\mathcal{R}_n}(C^{\mathrm{Bayes}}) = \Big\{B_{1,n}\frac{1}{k} + B_{2,n}\Big(\frac{k}{n}\Big)^{4/d}\Big\}\{1 + o(1)\} + o(t_n),$$

uniformly for $k \in K_{\beta,0}$. Observe that, by assumption **(B.2)**,

$$\frac{B_{1,n}}{k} \le \frac{1}{4k}\sup_{x_0\in\mathcal{S}}\Big\{\frac{1}{\|\dot{\eta}(x_0)\|}\Big\}\int_{\mathcal{S}\cap\mathcal{R}_n}\bar{f}(x_0)\,d\mathrm{Vol}^{d-1}(x_0) = O\Big(\frac{1}{k}\Big),$$

uniformly for $k \in K_{\beta,0}$. Furthermore, using assumption **(B.4)(ρ)**, we see that

$$\begin{aligned}
B_{2,n}\Big(\frac{k}{n}\Big)^{4/d} &= t_n^{\rho/(\rho+d)}\int_{\mathcal{S}\cap\mathcal{R}_n} t_n^{4/d-\rho/(\rho+d)}\frac{1}{\log^8(n/k)}\frac{\bar{f}(x_0)^{1-4/d}}{\|\dot{\eta}(x_0)\|}a(x_0)^2\,d\mathrm{Vol}^{d-1}(x_0)\\
&\le \sup_{x_0\in\mathcal{S}}\Big\{\frac{a(x_0)^2}{\|\dot{\eta}(x_0)\|}\Big\}t_n^{\rho/(\rho+d)}\int_{\mathcal{S}\cap\mathcal{R}_n} t_n^{4/d-\frac{\rho}{\rho+d}}\frac{1}{\log^8(n/k)}\bar{f}(x_0)^{1-4/d}\,d\mathrm{Vol}^{d-1}(x_0)\\
&\le \sup_{x_0\in\mathcal{S}}\Big\{\frac{a(x_0)^2}{\|\dot{\eta}(x_0)\|}\Big\}\frac{t_n^{\rho/(\rho+d)}}{\log^8(n/k)}\int_{\mathcal{S}\cap\mathcal{R}_n}\bar{f}(x_0)^{d/(\rho+d)}\,d\mathrm{Vol}^{d-1}(x_0).\\
&= O((k/n)^{\rho/(\rho+d)}\log^{2d\rho/(\rho+d)-8}(n/k)),
\end{aligned}$$

uniformly for $k \in K_{\beta,0}$. Finally, by the moment assumption in **(B.4)(ρ)** and Hölder's inequality, observe that

$$\begin{aligned}
&R_{\mathbb{R}^d\setminus\mathcal{R}_n}(\hat{C}_n^{\mathrm{knn}}) - R_{\mathbb{R}^d\setminus\mathcal{R}_n}(C^{\mathrm{Bayes}}) \le \int_{\mathcal{R}_n^c}\bar{f}(x)\,dx \le t_n^{\rho/(\rho+d)}\int_{\mathcal{R}_n^c}\bar{f}(x)^{1-\rho/(\rho+d)}\,dx\\
&\le t_n^{\rho/(\rho+d)}\Big\{\int_{\mathcal{R}_n^c}(1+\|x\|^\rho)\bar{f}(x)\,dx\Big\}^{1-\rho/(\rho+d)}\Big\{\int_{\mathcal{R}_n^c}\frac{1}{(1+\|x\|^\rho)^{(\rho+d)/\rho}}\,dx\Big\}^{\rho/(\rho+d)}\\
&= o(t_n^{\rho/(\rho+d)})
\end{aligned}$$

as $n \to \infty$, uniformly for $k \in K_{\beta,0}$. □

3.6.2 Tail adaptive results

Proof of Theorem 3.4. Recall that

$$k_{\mathrm{O}}(x) := \max\big[1, \min\big\{\big\lfloor B\big\{\bar{f}(x)n\big\}^{4/(d+4)}\big\rfloor, \lfloor n^{1-\beta}\rfloor\big\}\big].$$

Let $\mathcal{R}_n' := \{x : \bar{f}(x) \ge \frac{\log^{2(d+4)} n}{n}\} \subseteq \big\{x : \bar{f}(x) \ge \frac{k_{\mathrm{O}}(x)}{n}\log^{2d}(\frac{n}{k_{\mathrm{O}}(x)})\big\}$. Moreover, for $x \in \mathcal{R}_n'$, we have $k_{\mathrm{O}}(x) \in [B\log^8 n, n^{1-\beta}]$, we suppose, therefore, that n is large enough that $B\log^8 n > \log^4 n$. Observe that we can replace $\mathcal{R}_n$ in Theorem 3.6 with any compact subset $\mathcal{R} \subseteq \mathcal{R}_n$. Therefore, setting $\mathcal{R} = \mathcal{R}_n'$, $k_{\mathrm{L}} = k_{\mathrm{O}}$, and using also the fact

that $\bar{f}$ is continuous, we deduce from Theorem 3.6 that

$$R_{\mathcal{R}'_n}(\hat{C}_n^{k_{\mathrm{O}}\mathrm{nn}}) - R_{\mathcal{R}'_n}(C^{\mathrm{Bayes}}) = B_3 n^{-4/(d+4)}\{1 + o(1)\} \tag{3.39}$$

as $n \to \infty$.

It remains to bound the risk on $\mathbb{R}^d \setminus \mathcal{R}'_n$. *To prove Part (i)*: Fix $\alpha \in \big(0, 1 - \frac{4(\rho+d)}{(d+4)\rho}\big)$. Let $t'_n := \frac{\log^{2(d+4)} n}{n}$. By Markov's inequality and Hölder's inequality, observe that

$$\begin{aligned}
R_{\mathbb{R}^d\setminus\mathcal{R}'_n}(\hat{C}_n^{k_{\mathrm{O}}\mathrm{nn}}) - R_{\mathbb{R}^d\setminus\mathcal{R}'_n}(C^{\mathrm{Bayes}}) &\le {t'_n}^{\frac{4}{(d+4)(1-\alpha)}} \int_{\mathbb{R}^d\setminus\mathcal{R}'_n} \bar{f}(x)^{1-\frac{4}{d(1-\alpha)}}\, dx \\
&\le {t'_n}^{\frac{4}{(d+4)(1-\alpha)}} \Big\{\int_{\mathbb{R}^d\setminus\mathcal{R}'_n} (1 + \|x\|^\rho)\bar{f}(x)\, dx\Big\}^{1-\frac{4}{(d+4)(1-\alpha)}} \\
&\qquad \Big\{\int_{\mathbb{R}^d\setminus\mathcal{R}'_n} \frac{1}{(1+\|x\|^\rho)^{(d+4)(1-\alpha)/4-1}}\, dx\Big\}^{\frac{4}{(d+4)(1-\alpha)}} \\
&= o\Big(\Big(\frac{1}{n}\Big)^{4/(d+4)}\Big)
\end{aligned}$$

as $n \to \infty$, since $\rho\{(d+4)(1-\alpha)/4 - 1\} > d$. It remains to show B_3 is finite, this can be done in the same way as in the proof of Theorem 3.2, part (i).

To prove Part (ii): By the same argument as that in Theorem 3.2, part (ii), we have that

$$B_{3,n}\Big(\frac{1}{n}\Big)^{4/(d+4)} = O\Big(\Big(\frac{1}{n}\Big)^{\frac{\rho}{\rho+d}} \log^{2(d+4)\rho/(\rho+d)-8} n\Big).$$

Now, for the region $\mathbb{R}^d \setminus \mathcal{R}'_n$, observe that

$$\begin{aligned}
R_{\mathbb{R}^d\setminus\mathcal{R}'_n}(\hat{C}_n^{k\mathrm{nn}}) - R_{\mathbb{R}^d\setminus\mathcal{R}'_n}(C^{\mathrm{Bayes}}) &\le {t'_n}^{\rho/(\rho+d)} \int_{\mathbb{R}^d\setminus\mathcal{R}'_n} \bar{f}(x)^{1-\rho/(\rho+d)}\, dx \\
\le {t'_n}^{\rho/(\rho+d)} \Big\{\int_{\mathbb{R}^d\setminus\mathcal{R}'_n} (1+\|x\|^\rho)\bar{f}(x)\, dx\Big\}^{\frac{d}{\rho+d}} &\Big\{\int_{\mathbb{R}^d\setminus\mathcal{R}'_n} \frac{1}{(1+\|x\|^\rho)^{(\rho+d)/\rho}}\, dx\Big\}^{\frac{\rho}{\rho+d}} \\
&= o\Big(\Big(\frac{1}{n}\Big)^{\frac{\rho}{\rho+d}} \log^{2(d+4)\rho/(\rho+d)} n\Big)
\end{aligned}$$

as $n \to \infty$. □

Proof of Theorem 3.5. We prove parts (i) and (ii) of the theorem simultaneously, by appealing to the corresponding arguments in the proof of Theorem 3.4. First, we introduce the following class of functions: for $0 < c \le C$ and $n \in \mathbb{N}$, let

$$\mathcal{G}_{n,c,C} := \Big\{g : \mathbb{R}^d \to \mathbb{R} : g \ \text{ continuous}, c \le \inf_{x\in\mathcal{R}'_n} \frac{\bar{f}(x)}{g(x)} \le \sup_{x\in\mathcal{R}'_n} \frac{\bar{f}(x)}{g(x)} \le C\Big\}.$$

Now, for $g \in \mathcal{G}_{n,c,C}$, let

$$k_g(x) := \max[1, \min\{\lfloor B\{g(x)n\}^{4/(d+4)}\rfloor, \lfloor n^{1-\beta}\rfloor\}]. \tag{3.40}$$

Write

$$R(\hat{C}_n^{k_g\mathrm{nn}}) - R(C^{\mathrm{Bayes}}) \\ = R_{\mathcal{R}'_n}(\hat{C}_n^{k_g\mathrm{nn}}) - R_{\mathcal{R}'_n}(C^{\mathrm{Bayes}}) + R_{\mathbb{R}^d\setminus\mathcal{R}'_n}(\hat{C}_n^{k_g\mathrm{nn}}) - R_{\mathbb{R}^d\setminus\mathcal{R}'_n}(C^{\mathrm{Bayes}}).$$

To deal with the first term, we note that, for all sufficiently large n, $\mathcal{R}'_n \subseteq \{\bar{f}(x) \geq \frac{k_g(x)}{n}\log^{2d}(n/k_g(x))\}$ and $k_g \in K_\beta$, uniformly for $g \in \mathcal{G}_{n,c,C}$. We can therefore apply Theorem 3.6 (similarly to the application in the proof of Theorem 3.4) to conclude that

$$R_{\mathcal{R}'_n}(\hat{C}_n^{k_g\mathrm{nn}}) - R_{\mathcal{R}'_n}(C^{\mathrm{Bayes}}) = \Big\{\frac{1}{B}\int_{\mathcal{S}'_n}\frac{\bar{f}(x_0)^{d/(d+4)}}{4\|\dot{\eta}(x_0)\|}\Big(\frac{\bar{f}(x_0)}{g(x_0)}\Big)^{4/(d+4)}d\mathrm{Vol}^{d-1}(x_0) \\ + B^{4/d}\int_{\mathcal{S}'_n}\frac{\bar{f}(x_0)^{d/(d+4)}}{\|\dot{\eta}(x_0)\|}\Big(\frac{g(x_0)}{\bar{f}(x_0)}\Big)^{16/(d(d+4))}a(x_0)^2\,d\mathrm{Vol}^{d-1}(x_0)\Big\}n^{-4/(d+4)}\{1+o(1)\}, \tag{3.41}$$

uniformly for $g \in \mathcal{G}_{n,c,C}$. Moreover, for the tail region, we have that

$$R_{\mathbb{R}^d\setminus\mathcal{R}'_n}(\hat{C}_n^{k_g\mathrm{nn}}) - R_{\mathbb{R}^d\setminus\mathcal{R}'_n}(C^{\mathrm{Bayes}}) \leq \mathbb{P}(X \in \mathbb{R}^d \setminus \mathcal{R}'_n) = o\Big(\Big(\frac{1}{n}\Big)^{\frac{\rho}{\rho+d}}\log^{2(d+4)\rho/(\rho+d)}n\Big), \tag{3.42}$$

uniformly for $g \in \mathcal{G}_{n,c,C}$, due to the moment condition in **(B.4)(ρ)**.

Thus, by similar arguments to those in the proof of Theorem 3.4, parts (i) and (ii), we have that

(i) if $\rho > 4$, then

$$R(\hat{C}_n^{k_g\mathrm{nn}}) - R(C^{\mathrm{Bayes}}) = B_{4,g}n^{-4/(d+4)}\{1+o(1)\}, \tag{3.43}$$

uniformly for $g \in \mathcal{G}_{n,c,C}$, as $n \to \infty$, where

$$B_{4,g} := \frac{1}{B}\int_{\mathcal{S}'_n}\frac{\bar{f}(x_0)^{d/(d+4)}}{4\|\dot{\eta}(x_0)\|}\Big(\frac{\bar{f}(x_0)}{g(x_0)}\Big)^{4/(d+4)}d\mathrm{Vol}^{d-1}(x_0) \\ + B^{4/d}\int_{\mathcal{S}'_n}\frac{\bar{f}(x_0)^{d/(d+4)}}{\|\dot{\eta}(x_0)\|}\Big(\frac{g(x_0)}{\bar{f}(x_0)}\Big)^{16/(d(d+4))}a(x_0)^2\,d\mathrm{Vol}^{d-1}(x_0);$$

(ii) if $\rho \leq 4$, then

$$R(\hat{C}_n^{k_g\mathrm{nn}}) - R(C^{\mathrm{Bayes}}) = o(n^{-\rho/(\rho+d)}\log^{2(d+4)\rho/(\rho+d)}n),$$

uniformly for $g \in \mathcal{G}_{n,c,C}$, as $n \to \infty$.

Now, we show that $\hat{f}_m \in \mathcal{G}_{n,c,C}$ with high probability, for $m > m_0 n^{2+d/2}$ and n

large. First observe that, for all $x \in \mathcal{R}'_n$,

$$\Big|\frac{\hat{f}_m(x)}{\bar{f}(x)} - 1\Big| \leq \frac{n}{\log^{2(d+4)} n}|\hat{f}_m(x) - \bar{f}(x)| \leq \frac{n}{\log^{2(d+4)} n}\|\hat{f}_m - \bar{f}\|_\infty.$$

Write

$$\|\hat{f}_m - \bar{f}\|_\infty \leq \|\hat{f}_m - \mathbb{E}\hat{f}_m\|_\infty + \|\mathbb{E}\hat{f}_m - \bar{f}\|_\infty. \tag{3.44}$$

To bound the first term in (3.44), by Giné and Guillou (2002, Theorem 2.1), there exists $L > 0$, such that

$$\mathbb{P}(\|\hat{f}_m - \mathbb{E}\hat{f}_m\|_\infty \geq s m^{-2/(d+4)}) \leq L\exp\Big(\frac{-A^d s^2}{4L^2\|\bar{f}\|_\infty R(K)}\Big), \tag{3.45}$$

for all $2CA^{-d/2}\|\bar{f}\|_\infty^{1/2}R(K)^{1/2}\log^{1/2}\big(\frac{\|K\|_\infty m^{d/(2(d+4))}}{\|\bar{f}\|_\infty^{1/2}A^{d/2}R(K)^{1/2}}\big) \leq s \leq \frac{L\|\bar{f}\|_\infty R(K)m^{2/(d+4)}}{\|K\|_\infty}$, where $R(K) := \int_{\mathbb{R}^d}\|x\|^2K(x)\,dx$. Let

$$s_0 := \frac{2\max\{C, L\}\|\bar{f}\|_\infty^{1/2}R(K)^{1/2}}{A^{d/2}}\log^{2(d+4)} n.$$

Then, by applying the bound in (3.45) with $s = s_0$, for all $m_0 > \Big(\frac{8\max\{C,L\}\|\bar{f}\|_\infty^{1/2}R(K)^{1/2}}{A^{d/2}}\Big)^{2+d/2}$, we have that, for large n,

$$\begin{aligned}
\mathbb{P}\Big\{\|\hat{f}_m - \mathbb{E}\hat{f}_m\|_\infty \geq \frac{\log^{2(d+4)} n}{4n}\Big\} &\\
&\leq \mathbb{P}\Big\{\|\hat{f}_m - \mathbb{E}\hat{f}_m\|_\infty \geq \frac{2\max\{C, L\}\|\bar{f}\|_\infty^{1/2}R(K)^{1/2}\log^{2(d+4)} n}{A^{d/2}m_0^{2/(d+4)}n}\Big\}\\
&\leq \mathbb{P}\Big\{\|\hat{f}_m - \mathbb{E}\hat{f}_m\|_\infty \geq \frac{2\max\{C, L\}\|\bar{f}\|_\infty^{1/2}R(K)^{1/2}\log^{2(d+4)} n}{A^{d/2}m^{2/(d+4)}}\Big\}\\
&\leq L\exp\Big(-\log^{4(d+4)} n\Big) = O(n^{-M}),
\end{aligned}$$

for all $M > 0$. For the second term in (3.44), by a Taylor expansion, we have that, for all m_0, n sufficiently large,

$$\|\mathbb{E}\hat{f}_m - \bar{f}\|_\infty \leq \frac{A^2m^{-2/(d+4)}\sup_{x\in\mathbb{R}^d}\|\ddot{\bar{f}}(x)\|_{\text{op}}\int_{\mathbb{R}^d}K(z)\|z\|^2\,dz}{2} \leq \frac{\log^{2(d+4)} n}{4n}.$$

It follows that $\sup_{m>m_0n^{2+d/2}}\mathbb{P}\{\hat{f}_m \notin \mathcal{G}_{n,1/2,3/2}\} = O(n^{-M})$, for each $M > 0$, as $n \to \infty$.

To conclude the proof, let $x^{m\prime} := (x_{n+1}\dots x_{n+m})$ and let $\tilde{f}_m(x) = \tilde{f}_{m,h}(x, x^{m\prime}) :=$

$\frac{1}{mh^d}\sum_{j=1}^m K\big(\frac{x-x_{n+j}}{h}\big)$. For $\rho > 4$, we have that

$$
\begin{aligned}
&|R(\hat{C}_n^{k_{\mathrm{SS}}\mathrm{nn}}) - R(C^{\mathrm{Bayes}}) - B_4 n^{-4/(d+4)}| \\
&\le \int_{\{x^{m\prime}:\tilde{f}_m \in \mathcal{G}_{n,1/2,3/2}\}} |R(\hat{C}_n^{k_{\tilde{f}_m}\mathrm{nn}}) - R(C^{\mathrm{Bayes}}) - B_{4,\tilde{f}_m} n^{-4/(d+4)}|\, dP_X^{\otimes m}(x_{n+1},\dots,x_{n+m}) \\
&\qquad\qquad + \mathbb{P}\{\hat{f}_m \notin \mathcal{G}_{n,1/2,3/2}\} \\
&\le \sup_{g\in\mathcal{G}_{n,1/2,3/2}} |R(\hat{C}_n^{k_g\mathrm{nn}}) - R(C^{\mathrm{Bayes}}) - B_{4,g} n^{-4/(d+4)}| + O(n^{-M}) = o(n^{-4/(d+4)}),
\end{aligned}
$$

by (3.43). Similarly, for $\rho \le 4$,

$$
\begin{aligned}
&|R(\hat{C}_n^{k_{\mathrm{SS}}\mathrm{nn}}) - R(C^{\mathrm{Bayes}})| \\
&\qquad \le \int_{\{x^{m\prime}:\tilde{f}_m \in \mathcal{G}_{n,1/2,3/2}\}} |R(\hat{C}_n^{k_{\tilde{f}_m}\mathrm{nn}}) - R(C^{\mathrm{Bayes}})|\, dP_X^{\otimes m}(x_{n+1},\dots,x_{n+m}) \\
&\qquad\qquad + \mathbb{P}\{\hat{f}_m \notin \mathcal{G}_{n,1/2,3/2}\} \\
&\qquad \le \sup_{g\in\mathcal{G}_{n,1/2,3/2}} |R(\hat{C}_n^{k_g\mathrm{nn}}) - R(C^{\mathrm{Bayes}})| + O(n^{-M}) = o(n^{-\rho/(\rho+d)} \log^{2(d+4)\rho/(\rho+d)} n).
\end{aligned}
$$

□

Bibliography

Ailon, N. and Chazelle, B. (2006). Approximate nearest neighbours and the fast Johnson–Lindenstrauss transform. *Proceedings of the Symposium on Theory of Computing*.

Audibert, J.-Y. and Tsybakov, A. B. (2007). Fast learning rates for plug-in classifiers. *Ann. Statist.*, **35**, 608–633.

Azizyan, M., Singh, A. and Wasserman, L. (2013). Density-sensitive semisupervised inference. *Ann. Statist.*, **41**, 751–771.

Bartlett, P. and Traskin, M. (2007). AdaBoost is consistent. *J. Mach. Learn. Res.*, **8**, 2347–2368.

Biau, G., Cérou, F. and Guyader, A. (2010). On the rate of convergence of the bagged nearest neighbor estimate. *J. Mach. Learn. Res.*, **11**, 687–712.

Biau, G., Devroye, L. and Lugosi, G. (2008). Consistency of random forests and other averaging classifiers. *J. Mach. Learn. Res.*, **9**, 2015–2033.

Bickel, P. J. and Levina, E. (2004). Some theory for Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are more variables than observations. *Bernoulli*, **10**, 989–1010.

Bijral, A. S., Ratliff, N. and Srebro, N. (2012). Semi-supervised learning with density based distances. *ArXiv e-prints*, 1202.3702.

Blaser, R. and Fryzlewicz, P. (2015). Random rotation ensembles. *Preprint*. Available from `http://stats.lse.ac.uk/fryzlewicz/articles.html`.

Boucheron, S., Bousquet, O. and Lugosi, G. (2005). Theory of classification: A survey of some recent advances. *ESAIM Probab. Stat.*, **9**, 323375.

Breiman, L. (1996). Bagging Predictors. *Machine Learning*, **24**, 123–140.

Breiman, L. (2001). Random Forests. *Machine Learning*, **45**, 5–32.

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

Cai, T. and Liu, W. (2011). A direct estimation approach to sparse linear discriminant analysis. *J. Amer. Statist. Assoc.*, **106**, 1566-1577.

Cannings, T. I. and Samworth, R. J. (2015a). Random projection ensemble classification. *ArXiv e-prints*, 1504.04595.

Cannings, T. I. and Samworth, R. J. (2015b). `RPEnsemble`: Random projection ensemble classification. `R` package version 0.2, `http://cran.r-project.org/web/packages/RPEnsemble/`.

Chapelle, O., Zien, A. and Schölkopf, B. (Eds.) (2006). *Semi-supervised Learning*. MIT Press, Cambridge MA.

Chanda, K. C. and Ruymgaart, F. (1989). Asymptotic estimate of probability of misclassification for discriminant rules based on density estimates. *Statist Probab. Lett.*, **8**, 81–88.

Chen, Y. and Samworth, R. J. (2013) Smoothed log-concave maximum likelihood estimation with applications. *Statist. Sinica*, **23**, 1373–1398.

Clemmensen, L., Hastie, T., Witten, D. and Ersbøll, B. (2011). Sparse discriminant analysis. *Technometrics*, **53**, 406-413.

Cook, R. D. (1998). *Regression Graphics: Ideas for Studying Regressions through Graphics*. Wiley, New York.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, **20**, 273–297.

Cover, T. M. and Hart, P. E. (1967). Nearest neighbour pattern classification. *IEEE Trans. Inf. Th.*, **13**, 21–27.

Dasgupta, S. (1999). Learning mixtures of Gaussians. *Proc. 40th Annual Symposium on Foundations of Computer Science*, 634–644.

Dasgupta, S. and Gupta, A. (2002). An elementary proof of the Johnson–Lindenstrauss Lemma. *Rand. Struct. Alg.*, **22**, 60–65.

Devroye, L. P. and Wagner, T. J. (1976). A distribution-free performance bound in error estimation. *IEEE Trans. Info. Th.*, **22**, 586–587.

Devroye, L. P. and Wagner, T. J. (1979). Distribution-free inequalities for the deleted and hold-out error estimates. *IEEE Trans. Info. Th.*, **25**, 202–207.

Devroye, L., Györfi, L. and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*. Springer, New York.

Duda, R. O., Hart, P. E. and Stork, D. G. (2000). *Pattern Classification, 2nd ed.* Wiley, New York.

Dudley, R. M. (1999) *Uniform Central Limit Theorems*. Cambridge University Press, Cambridge.

Duong, T. (2015). `ks`: Kernel smoothing. `R` package version 1.9.4, `https://cran.r-project.org/web/packages/ks`.

Durrant, R. J. and Kaban, A. (2013). Sharp generalization error bounds for randomly-projected classifiers. *Proc. J. Mach. Learn. Res.*, **28**, 693–701.

Durrant, R. J. and Kaban, A. (2014). Random projections as regularizers: learning a linear discriminant from fewer observations than dimensions. *Machine Learning*, to appear.

Esseen, C.-G. (1945). Fourier analysis of distribution functions. A mathematical study of the Laplace–Gaussian law. *Acta Mathematica*, **77**, 1–125.

Fan, J. and Fan, Y. (2008). High-dimensional classification using features annealed independence rules. *Ann. Statist.*, **36**, 2605–2637.

Fan, J., Fan, Y. and Wu, Y. (2010). High-dimensional classification. In *High-dimensional Data Analysis*. (Eds. Cai, T. T. and Shen, X.), pp. 3-37. World Scientific, New Jersey.

Fan, J., Feng, Y., Jiang, J. and Tong, X. (2015). Feature augmentation via nonparametrics and selection (FANS) in high dimensional classification. *J. Amer. Statist. Assoc.*, to appear.

Fan, J., Feng, Y. and Tong, X. (2012). A ROAD to classification in high dimensional space: the regularized optimal affine discriminant. *J. Roy. Statist. Soc., Ser. B.*, **72**, 745–771.

Feller, W. (1957). *An Introduction to Probability Theory and Its Applications vol. 2,* 2nd ed. Wiley, New York.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, **7**, 179–188.

Fix, E. and Hodges, J. L. (1951). Discriminatory analysis – nonparametric discrimination: Consistency properties. Technical Report 4, USAF School of Aviation Medicine, Randolph Field, Texas.

Fix, E. and Hodges, J. L. (1989). Discriminatory analysis – nonparametric discrimination: Consistency properties. *Internat. Statist. Rev.*, **57**, 238–247.

Freund, Y. and Schapire, R. E. (1997). A decision-theoretic generalisation of on-line learning and an application to boosting. *J. Computer Systems Sci.*, **55**, 119-139.

Freund, Y. and Schapire, R. E. (2012). *Boosting: Foundations and Algorithms.* The MIT press, Cambridge MA.

Friedman, J. (1989). Regularised discriminant analysis. *J. Amer. Statist. Assoc.*, **84**, 165–175.

Friedman, J. (1994). Flexible metric nearest neighbor classification. *Technical report.*

Gadat, S., Klein, T. and Marteau, C. (2014). Classification with the nearest neighbour rule in general finite dimensional spaces. *ArXiv e-prints*, 1411.0894.

Giné, E. and Guillou, A. (2002). Rates of strong uniform consistency for multivariate kernel density estimators. *Ann. Inst. H. Poincaré Probab. Statist.*, **38**, 907-921.

Gnedenko, B. V. and Kolmogorov, A. N. (1954). *Limit Distributions for Sums of Independent Random Variables*, Addison-Wesley, Cambridge MA.

Gordon, A. D. (1999). *Classification, 2nd ed.* Chapman & Hall, London.

Gray, A. (2004). *Tubes, 2nd ed.* Progress in Mathematics 221. Birkhäuser, Basel.

Guo, Y., Hastie, T. and Tibshirani, R. (2007). Regularised linear discriminant analysis and its application in microarrays. *Biostatistics*, **8**, 86–100.

Hall, P. and Kang, K.-H. (2005). Bandwidth choice for nonparametric classification. *Ann. Statist.*, **33**, 284–306.

Hall, P., Park, B. U. and Samworth, R. J. (2008). Choice of neighbour order in nearest-neighbour classification. *Ann. of Statist.*, **36**, 2135–2152.

Hall, P. and Samworth, R. J. (2005). Properties of bagged nearest neighbour classifiers. *J. Roy. Statist. Soc., Ser. B.*, **67**, 363–379.

Hand, D (1981). *Classification and Discrimination.* Wiley Series in Probability and Statistics - Applied Probability and Statistics Section, Wiley, New York.

Hao, N., Dong, B. and Fan, J. (2015). Sparsifying the Fisher linear discriminant by rotation. *J. Roy. Statist. Soc., Ser. B.*, to appear.

Hastie, T., Buja, A. and Tibshirani, R. (1995). Penalized discriminant analysis. *Ann. Statist.*, **23**, 73–102.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning: The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer Series in Statistics (2nd ed.). Springer, New York.

Hastie, T., Tibshirani, R., Narisimhan, B. and Chu, G. (2015). `pamr`: Pam: prediction analysis for microarrays. `R` package version 1.55, `http://cran.r-project.org/web/packages/pamr/`.

Hilbert, M. and López, P. (2011). The world's technological capacity to store, communicate, and compute information. *Science*, **332**, 60-65.

Islam, M. S. and McLeod, A. I. (2015). *ascrda*: Augmented SCRDA. `R` package version 1.15, `http://cran.r-project.org/web/packages/ascrda/`

Jones, M. C., Marron, J. S. and Sheather, S. J. (1996). A brief survey of bandwidth selection for density estimation. *J. Amer. Statist. Assoc.*, **91**, 401-407.

Karatzoglou, A., Smola A. and Hornik, K. (2015). `kernlab`: Kernel-based Machine Learning Lab. `R` package version 0.9-20, `http://cran.r-project.org/web/packages/kernlab/`

Kim, A. K.-H. and Samworth, R. J. (2014). Global rates of convergence in log-concave density estimation. *arXiv e-prints*, `1404.2298`.

Laber, E. and Murphy, S. (2011). Adaptive confidence intervals for the test error in classification. *J. Amer. Statist. Assoc.*, **106**, 904-913.

Lee, K.-Y., Li, B. and Chiaromonte, F. (2013). A general theory for nonlinear sufficient dimension reduction: formulation and estimation. *Ann. Statist.*, **41**, 221–249.

Li, K.-C. (1991). Sliced inverse regression for dimension reduction. *J. Amer. Statist. Assoc.*, **86**, 316–342.

Liaw, A. and Wiener, M. (2014). `randomForest`: Breiman and Cutler's random forests for classification and regression. `R` package version 4.6-10, `http://cran.r-project.org/web/packages/randomForest/`.

Liang, F., Mukherjee, S. and West, M. (2007). The use of unlabeled data in predictive modeling. *Statist. Sci.*, **22**, 189–205.

Lopes, M. (2013). The convergence rate of majority vote under exchangeability. Technical report. http://arxiv.org/abs/1303.0727

Lopes, M., Jacob, L. and Wainwright, M. J. (2011). A more powerful two-sample test in high dimensions using random projection. *Advances in Neural Information Processing Systems (NIPS)*.

Mammen, E. and Tsybakov, A. B. (1999). Smooth discriminant analysis. *Ann. Statist.*, **27**, 1808–1829.

Marzetta, T., Tucci, G. and Simon, S. (2011). A random matrix-theoretic approach to handling singular covariance estimates. *IEEE Trans. Inf. Th.*, **57**, 6256–6271.

Mayer-Schönberger, V. and Cukier, K. (2015). *Big Data: A Revolution That Will Transform How We Live, Work and Think.* Houghton Mifflin Harcourt, New York.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *J. Roy. Statist. Soc., Ser. B*, **72**, 417–473.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel A., Leisch, F., Chang, C.-C. and Lin, C.-C. (2015). e1071: Misc Functions of the Department of Statistics (e1071), TU Wien. R package version 1.6-4, http://cran.r-project.org/web/packages/e1071/.

McWilliams, B., Heinze, C., Meinshausen, N., Krummenacher, G. and Vanchinathan, H. P. (2014). LOCO: distributing ridge regression with random projections. *arXiv e-prints*, 1406.3469v2.

Okamoto, M. (1963). An asymptotic expansion for the distribution of the linear discriminant function. *Ann. Math. Statist.*, **34**, 1286–1301.

Owen, A. (1984). A neighbourhood-based classifier for LANDSAT data. *The Canadian J. Statist.*, **12**, 191–200.

Parzen, E. (1962). On estimation of a probability density function and mode. *The Ann. Math. Statist.* **33**, 1065–1076.

Petrov, V. V. (1975). *Sums of Independent Random Variables.* Springer-Verlag, Berlin.

Rigollet, P. (2007). Generalization error bounds in semi-supervised classification under the cluster assumption. *J. Mach. Learn. Res.*, **8**, 1369–1392.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *Ann. Math. Statist.*, **27**, 832–837.

Samworth, R. J. (2012a). Optimal weighted nearest neighbour classifiers. *Ann. Statist.*, **40**, 2733–2763.

Samworth, R. J. (2012b). Online supplement to 'Optimal weighted nearest neighbour classifiers'.

Samworth, R. J. and Wand, M. P. (2010). Asymptotics and optimal bandwidth selection for highest density region estimation. *Ann. Statist.*, **38**, 1767-1792.

Schapire, R. E. (2013). Explaining AdaBoost. In *Empirical Inference* (Eds. Schölkopf, B., Luo, Z. and Vovk, V.), pp. 37–57. Springer, Berlin.

Shah, R. D. and Samworth, R. J. (2013). Variable selection with error control: another look at stability selection. *J. Roy. Statist. Soc., Ser. B*, **75**, 55–80.

Silverman, B. (1986). *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London.

Stone, C. J. (1977). Consistent nonparametric regression. *Ann. Statist.*, **5**, 595–620.

Tibshirani, R., Hastie, T., Narisimhan, B. and Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Procedures of the Natural Academy of Science, USA*, **99**, 6567–6572.

Tibshirani, R., Hastie, T., Narisimhan, B. and Chu, G. (2003). Class prediction by nearest shrunken centroids, with applications to DNA microarrays. *Statist. Science*, **18**, 104–117.

Tsybakov, A. B. (2004). Optimal aggregation of classifiers in statistical learning. *Ann. Statist.*, **32**, 135–166.

Vapnik,V. N. (2000). *The Nature of Statistical Learning Theory, 2nd ed.* Springer, New York.

Vapnik, V. and Chervonenkis, A. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Th. Probab. Appli.*, **16**, 264-380.

Vapnik, V. and Chervonenkis, A. (1974a). Ordered risk minimisation. I. *Auto. and Remote Control*, **35**, 1226-1235.

Vapnik, V. and Chervonenkis, A. (1974b). Ordered risk minimisation. II. *Auto. and Remote Control*, **35**, 1403-1412.

Vapnik, V. and Chervonenkis, A. (1974c). *Theory of Pattern Recognition* Nauka, Moscow. (in Russian).

Vershynin, R. (2012). Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing* (Eds. Eldar, Y. C. and Kutyniok, G.), pp. 210–268. Cambridge University Press, Cambridge.

Walther, G. (2009). Inference and modeling with log-concave distributions. *Statist. Science*, **24**, 319-327.

Williams, C. K. I. and Barber, D. (1998). Bayesian classification with Gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **20**, 1342–1351.

Witten, D. (2011). `penalizedLDA`: Penalized classification using Fisher's linear discriminant. `R` package version 1.0, `http://cran.r-project.org/web/packages/penalizedLDA/`.

Witten, D. M. and Tibshirani, R. (2011). Penalized classification using Fisher's linear discriminant. *J. Roy. Statist. Soc., Ser. B.*, **73**, 753–772.

Xue, J.-H. and Hall, P. (2015). Why does rebalancing class-unbalanced data improve AUC for linear discriminant analysis? *IEEE Trans. Pat. Anal. Mach. Intel.*, **37**, 1109-1112.