Nearest neighbour classification in the tails of a distribution

Timothy I. Cannings
Statistical Laboratory, University of Cambridge

Abstract

We discuss the theoretical properties of weighted nearest neighbour classifiers. Samworth (2012) derived an asymptotic expansion for the excess risk of a weighted nearest neighbour classifier over any compact set when the feature vectors are d dimensional. It is shown that when $d = 1$, under suitable conditions on the tails of the distributions, the asymptotic expansion of the excess risk remains valid on the whole of $\mathbb{R}$. Without such conditions we see there may be an additional contribution to the excess risk of the same order arising from the tails of the marginal distribution. For $d \geq 2$ the problem is more difficult, and we discuss where the results in one dimension may help us with future work.

1 Introduction

The problem of classification can be thought of as the decision process used in assigning an object, or objects, to one of several groups. Think for example, of a doctor making a medical diagnosis, an email spam filter determining whether an email is genuine, an expert performing an authorship analysis, a machine deciding whether certain items pass a quality control test, or identification of an animal's species from an image. In this paper we focus on the supervised setting, that is, we know in advance the characteristics or features of a number of objects and the class to which each object belongs. This simple problem has led to a rich body of literature on different methods and the theory behind them. Notable early works include Fix & Hodges (1951) and Cover & Hart (1967); there are also many books on the subject (Devroye, Gyorfi & Lugosi, 1996; Hand, 1981).

1.1 Statistical setting

We work in the following setting: suppose we have independent, identically distributed pairs $(X, Y), (X_1, Y_1), \ldots, (X_n, Y_n)$ taking values in $\mathbb{R}^d \times \{1, 2\}$, with $\mathbb{P}(Y = 1) = p = 1 - \mathbb{P}(Y = 2)$ 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 the measures $\bar{P} := pP_1 + (1 - p)P_2$ and $P^\circ := pP_1 - (1 - p)P_2$. We are presented with the task of assigning a new object X to either class 1 or 2. A *classifier* is a Borel measurable function C from $\mathbb{R}^d$ to $\{1, 2\}$ with the understanding that C assigns a point $x \in \mathbb{R}^d$ to the class $C(x)$. Given a Borel measurable set $\mathcal{R} \subset \mathbb{R}^d$ we would like classifiers with a low misclassification rate, or risk over $\mathcal{R}$, that is

$$\mathcal{R}_{\mathcal{R}}(C) := \mathbb{P}\{C(X) \neq Y, X \in \mathcal{R}\}.$$

Observe that,

$$\mathbb{P}\{C(X) \neq Y, X \in \mathcal{R}\} = \int_{\mathcal{R}} p\mathbb{P}\{C(X) = 2\}P_1(dx) + \int_{\mathcal{R}} (1-p)\mathbb{P}\{C(X) = 1\}P_2(dx). \quad (1.1)$$

Definition Define the *Bayes classifier*

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

The following lemma can be found in Devroye et al. (1996, p. 10).

Lemma 1.1. *The Bayes classifier minimises* $\mathbb{P}\{C(X) \neq Y, X \in \mathcal{R}\}$.

Proof. Let C^{Bayes} be the Bayes classifier as above, and C be any other classifier. Then

$$\begin{aligned} \mathbb{P}\{C(X) = Y|X = x\} &= \mathbb{P}\{C(X) = 1, Y = 1|X = x\} + \mathbb{P}\{C(X) = 2, Y = 2|X = x\} \\ &= \mathbb{P}(Y = 1|X = x)\mathbb{1}_{\{C(x)=1\}} + \mathbb{P}(Y = 2|X = x)\mathbb{1}_{\{C(x)=2\}} \\ &= \eta(x)\mathbb{1}_{\{C(x)=1\}} + \{1 - \eta(x)\}\mathbb{1}_{\{C(x)=2\}}. \end{aligned}$$

It follows that

$$\begin{aligned} &\mathbb{P}\{C(X) \neq Y|X = x\} - \mathbb{P}\{C^{Bayes}(X) \neq Y|X = x\} \\ &\quad = \eta(x)(\mathbb{1}_{\{C^{Bayes}(x)=1\}} - \mathbb{1}_{\{C(x)=1\}}) + \{1 - \eta(x)\}(\mathbb{1}_{\{C^{Bayes}(x)=2\}} - \mathbb{1}_{\{C(x)=2\}}). \\ &\quad = \{2\eta(x) - 1\}(\mathbb{1}_{\{C^{Bayes}(x)=1\}} - \mathbb{1}_{\{C(x)=1\}}) \geq 0, \end{aligned}$$

by the definition of C^{Bayes}. The result follows by integrating over $\mathcal{R}$ with respect to x. □

In practice, the function η will be unknown. Nevertheless Lemma 1.1 gives us a benchmark to aim for: Can we find classifiers that perform as well as the Bayes classifier? With that in mind, the *excess risk* of a classifier C is defined to be $\mathcal{R}_{\mathcal{R}}(C) - \mathcal{R}_{\mathcal{R}}(C^{Bayes}) \geq 0$. Denote by $\hat{C}_n$ any classifier that depends on a training data set of size n. We will say $\hat{C}_n$ is *consistent* if its excess risk converges to zero as $n \to \infty$. Furthermore, for consistent classifiers we can then discuss the rate of convergence of the excess risk.

One might assume some parametric model for the conditional distributions P_1 and P_2, and then estimate the parameters using the training data. Methods of this type are widely used in practice and have been shown to work well in many situations. However, if the model is misspecified the resulting classifier can perform poorly (Devroye et al., 1996, p. 49). Alternative methods do not rely on a parametric model assumption, for example kernel-based classifiers (Fix & Hodges, 1951) and nearest neighbour methods (Cover & Hart, 1967). These nonparametric techniques can be shown to be consistent for large classes of distributions. Stone (1977) proved that under certain conditions on the choice of k (see later) the popular k-nearest neighbour classifier is consistent no matter what the marginal distributions. Hall & Kang (2005) and Samworth (2012) derived, under suitable regularity conditions, asymptotic expansions for the rate of convergence of the excess risk over fixed compact sets for kernel and weighted nearest neighbour classifiers respectively. It can be shown that these rates are minimax over certain classes (Marron, 1983; Audibert & Tsybakov, 2007).

For the remainder of this paper we will focus on nearest neighbour classifiers. The main purpose is to determine conditions on the tail behaviour of the measures P_1 and P_2 such that the asymptotic expansion for the excess risk of a weighted nearest neighbour classifier (derived by Samworth (2012)) remains valid over the whole of $\mathbb{R}^d$. Section 1.3 explores some examples for motivation and the main results are contained in Sections 2 and 3. We start by discussing the properties of nearest neighbour classifiers in more detail.

1.2 Nearest Neighbour Classifiers

Given a point $x \in \mathbb{R}^d$, let $(X_{(1)}, Y_{(1)}), (X_{(2)}, Y_{(2)}), \ldots, (X_{(n)}, Y_{(n)})$ be a reordering of the training data such that, $\|X_{(1)} - x\| \leq \|X_{(2)} - x\| \leq \cdots \leq \|X_{(n)} - x\|$, where $\|\cdot\|$ is some norm on $\mathbb{R}^d$ (for the purpose of this report we can think of $\|\cdot\|$ as the Euclidean norm). We can then define the regression estimate $\frac{1}{k}\sum_{i=1}^{k} \mathbb{1}_{\{Y_{(i)}=1\}}$ at x and the corresponding *k-nearest neighbour* classifier

$$\hat{C}_n^{knn}(x) = \begin{cases} 1 & \text{if } \frac{1}{k}\sum_{i=1}^{k} \mathbb{1}_{\{Y_{(i)}=1\}} \geq \frac{1}{2}; \\ 2 & \text{otherwise.} \end{cases} \tag{1.2}$$

This can be generalised by using a weighted sum of the nearest neighbours. Let $\mathbf{w}_n :=$

$\{w_{ni}\}_{i=1}^n$ be a weight vector with $\sum_{i=1}^n w_{ni} = 1$. Replacing the regression estimate above by the weighted sum $S_n(x) := \sum_{i=1}^n w_{ni} \mathbb{1}_{\{Y_{(i)}=1\}}$, we define the *weighted nearest neighbour* classifier

$$\hat{C}_n^{wnn}(x) = \begin{cases} 1 & \text{if } \sum_{i=1}^n w_{ni} \mathbb{1}_{\{Y_{(i)}=1\}} \geq \frac{1}{2}; \\ 2 & \text{otherwise.} \end{cases} \tag{1.3}$$

Stone (1977) proved the following powerful result; for a proof see Devroye et al. (1996, p. 98).

Theorem 1.2. *Suppose the the weights $\{w_{ni}\}_{i=1}^n$ satisfy the following conditions; (i) $\max_{i=1,\ldots,n} w_{ni} \to 0$ as $n \to \infty$, and (ii) $\sum_{i=1}^k w_{ni} \to 1$ as $n \to \infty$ for some $k := k_n$ such that $k_n \to \infty$ and $k_n/n \to 0$. Then the weighted nearest neighbour classifier is universally consistent, i.e. for any distributions P_1 and P_2 and any prior probability p,*

$$\mathcal{R}_{\mathbb{R}^d}(\hat{C}_n^{wnn}) - \mathcal{R}_{\mathbb{R}^d}(C^{Bayes}) \to 0.$$

Remarks Note that the conditions on the weights here include the unweighted k-nearest neighbour classifier provided $k := k_n$ is such that $k_n \to \infty$ and $k_n/n \to 0$.

1.2.1 Bagged Nearest Neighbour Classifiers

Consider the 1-nearest neighbour classifier; that is, the classifier that assigns a point to the same class as that of its nearest neighbour. This procedure can be shown to be inconsistent for many distributions. In other words, there exists $\epsilon > 0$ such that for all $n_0 > 1$ there exists $n > n_0$ with $\mathcal{R}_{\mathcal{R}}(\hat{C}_n^{1nn}) - \mathcal{R}_{\mathcal{R}}(C^{Bayes}) > \epsilon$. However, one can repeatedly use a 1-nearest neighbour classifier on bootstrap resamples of the training data and subsequently classify to the modal class arising from the resamples. This method is known as *bagging* (short for bootstrap aggregating). Hall & Samworth (2005) and Biau & Devroye (2010) proved analogous results to that of Stone (1977) and the large sample properties of this procedure were further studied by Biau, Cérou & Guyader (2010). Bagging can be regarded as a natural way to choose the weights for a weighted nearest neighbour classifier.

1.2.2 Optimal Weighted Nearest Neighbour Classifiers

Here we discuss the main theory in Samworth (2012). Firstly we require some more notation: let $B_\delta(x)$ denote the closed ball of radius δ centred at x for the norm $\|\cdot\|$, and let a_d denote the d-dimensional Lebesgue measure of the unit ball $B_1(x)$. For a smooth function $g : \mathbb{R}^d \to \mathbb{R}$ write $\dot{g}(x)$ for its gradient vector at x. For $\beta > 0$ let $W_{n,\beta}$ be the set of non-negative weight vectors $\mathbf{w}_n = \{w_{ni}\}_{i=1}^n$ satisfying

- $s_n^2 := \sum_{i=1}^n w_{ni}^2 \leq n^{-\beta}$;

- $t_n^2 := \left(\frac{\sum_{i=1}^n \alpha_i w_{ni}}{n^{2/d}} \right)^2 \leq n^{-\beta}$, where $\alpha_i = i^{(1+2/d)} - (i-1)^{(1+2/d)}$;
- $n^{2/d} \sum_{i=k_2+1}^n w_{ni} / \sum_{i=1}^n \alpha_i w_{ni} \leq 1/\log n$, where $k_2 = \lfloor n^{1-\beta} \rfloor$;
- $\sum_{i=k_2+1}^n w_{ni}^2 / \sum_{i=1}^n w_{ni}^2 \leq 1/\log n$;
- $\sum_{i=1}^n w_{ni}^3 / (\sum_{i=1}^n w_{ni}^2)^{3/2} \leq 1/\log n$.

We also make use of the following assumptions;

A. 1. *The set $\mathcal{R} \subset \mathbb{R}^d$ is a compact d-dimensional manifold with boundary $\partial\mathcal{R}$.*

A. 2. *The set $\mathcal{S} = \{x \in \mathcal{R} : \eta(x) = \frac{1}{2}\}$ is non-empty. There exists an open set U_0 that contains $\mathcal{S}$ and such that the restrictions of P_1 and P_2 to U_0 are absolutely continuous with respect to Lebesgue measure, with twice continuously differentiable Radon–Nikodym derivatives f_1 and f_2 respectively. Furthermore, η is continuous on U, where U is an open set containing $\mathcal{R}$.*

A. 3. *There exists $\rho > 0$ such that $\int_{\mathbb{R}^d} \|x\|^\rho d\bar{P}(x) < \infty$. Moreover, for sufficiently small $\delta > 0$, the ratio, $\bar{P}(B_\delta(x))/(a_d \delta^d)$ is bounded away from zero uniformly over $x \in \mathcal{R}$.*

A. 4. *For all $x \in \mathcal{S}$, we have $\dot{\eta}(x) \neq 0$, and for all $x \in \mathcal{S} \cap \partial\mathcal{R}$, we have $\dot{\partial\eta}(x) \neq 0$, where $\partial\eta$ denotes the restriction of η to $\partial\mathcal{R}$.*

The following theorem gives an asymptotic expansion of the excess risk under these assumptions.

Theorem 1.3. *(Samworth, 2012). Assume **(A. 1)**, **(A. 2)**, **(A. 3)**, and **(A. 4)**. Then for each $\beta \in (0, 1/2)$,*

$$\mathcal{R}_{\mathcal{R}}(\hat{C}_n^{wnn}) - \mathcal{R}_{\mathcal{R}}(C^{Bayes}) = \gamma_n(\mathbf{w}_n)\{1 + o(1)\}, \tag{1.4}$$

as $n \to \infty$, uniformly for $\mathbf{w_n} \in W_{n,\beta}$, where

$$\gamma_n(\mathbf{w}_n) = B_1 \sum_{i=1}^n w_{ni}^2 + B_2 \left(\sum_{i=1}^n \frac{\alpha_i w_{ni}}{n^{2/d}} \right)^2. \tag{1.5}$$

Explicit expressions for the constants B_1 and B_2 can be found in Samworth (2012).

We do not prove this here, indeed the proof is rather long. Instead we outline the main argument in two parts. Part one involves bounding the dominant contribution to the excess risk from points outside of a small tube around the boundary $\mathcal{S}$. One can show that the k_2 nearest neighbours to a point x are concentrated on a small ball centred at x with high probability. Therefore, when x is bounded away from $\mathcal{S}$, the k_2th nearest neighbour will be the *correct* side of the boundary. As a result the weighted nearest neighbour classifier classifies such points in the same way as the Bayes classifier with high probability. It is shown the excess risk arising

from this region is $\mathcal{O}(n^{-M})$ for all $M > 0$. For points near the boundary $\mathcal{S}$ we use a bias-variance decomposition argument to determine the dominant contribution to the risk. In the asymptotic expansion above the $\sum_{i=1}^{n} w_{ni}^2$ term arises from the variance, and the $\sum_{i=1}^{n} \frac{\alpha_i w_{ni}}{n^{2/d}}$ from the bias. Consequently there is a bias-variance trade-off in the choice of weights: with few positive weights the bias will be small and the variance large. Increasing the number of positive weights will reduce the variance, but only at the cost of increased bias.

More specifically Theorem 1.3 helps us to choose the weights in theory. We can minimise the right hand side of (1.4) over $\mathbf{w}_n \in W_{n,\beta}$ (or a subset thereof, for example restricting ourselves to unweighted classifiers) to obtain optimal weights. The optimal weighting scheme $\mathbf{w}_n^*$ is specified as follows: set $k^* = \lfloor B^* n^{\frac{4}{d+4}} \rfloor$ (an explicit expression for B^* is given in Samworth (2012)) and let

$$w_{ni}^* = \begin{cases} \frac{1}{k^*}\left[1 + \frac{d}{2} - \frac{d}{2k^{*2/d}}\{i^{1+2/d} - (i-1)^{1+2/d}\}\right] & \text{for } i = 1, 2, \ldots, k^*; \\ 0 & \text{for } i = k^*+1, \ldots, n. \end{cases} \tag{1.6}$$

A proportion $\mathcal{O}(n^{-d/(d+4)})$ of the optimal weights are positive, and under this scheme the rate of convergence of the excess risk to zero is $\mathcal{O}(n^{-4/(d+4)})$. Moreover the theory provides a simple, distribution independent, formula to go from an *optimal* unweighted nearest neighbour classifier to an optimal weighted one and give guaranteed improvements in the constant in the asymptotic expansion.

The assumption **(A. 4)** is closely related to the *margin condition*. Under this condition, and mild assumptions on the smoothness of the Bayes decision boundary, Mammen & Tsybakov (1999) and later Tsybakov (2005) investigate the rate of convergence of *plug-in* and *empirical risk minimisation* type classifiers.

Although the theory in Samworth (2012) is powerful, there is a drawback: the result relies on the compactness of $\mathcal{R}$. In practice, to guarantee this rate of convergence, we would need to specify a compact set in advance, and then ignore any observations we might observe outside this set. We might expect that if the tails of the distribution are light enough, then we would rarely observe such points, and the asymptotic expansion may still be valid.

1.3 Motivating Examples

We have seen that results on the rate of convergence of the excess risk are typically restricted to an expansion over a compact set. Moreover, under certain regularity conditions on the underlying distributions, the dominant contribution to the asymptotic expansion arises from a small interval around the boundary $\{x : \eta(x) = 1/2\}$. The following examples, which focus on the $d = 1$ case, provide some insight as to when the task of classification in the tails of the

distribution (i.e. outside of any fixed compact set) is difficult. By appealing to the theory in Samworth (2012), intuition suggests that classification in the tails will be hard in two situations. Firstly, when the conditional distributions are similar in the tails so that $\eta(x)$ is close to $1/2$ and secondly, when the tails are heavy. In each of the examples below we have (X, Y) taking values in $\mathbb{R} \times \{1, 2\}$ with densities f_1 and f_2 for P_1 and P_2 respectively.

1.3.1 Normal Location Model

Suppose (X, Y) has joint distribution specified by $p = 1/2$, $X|Y = 1 \sim \mathcal{N}(\mu, \sigma^2)$ and $X|Y = 2 \sim \mathcal{N}(-\mu, \sigma^2)$.

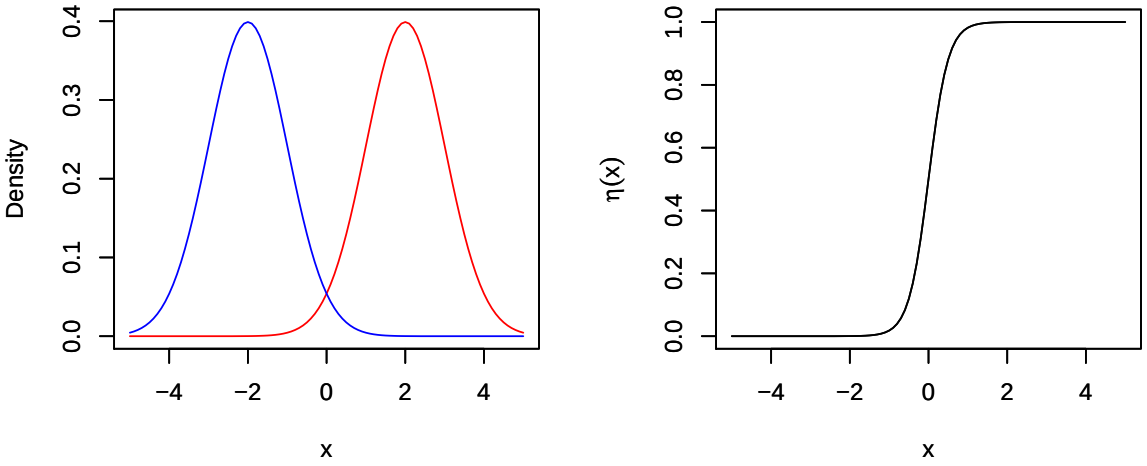

Figure 1: Left panel: The densities given $Y = 1$ (red) and $Y = 2$ (blue) for $\mu = 2$ and $\sigma = 1$. Right panel: The corresponding regression function.

Here we see a situation in which classification is 'easy', the two distributions are light-tailed and sufficiently different so that the regression function $\eta(x) \to 1$ as $x \to \infty$ and $\eta(x) \to 0$ as $x \to -\infty$.

1.3.2 Cauchy Location Model 1

Suppose (X, Y) is such that $p = 1/2$ and $X|Y = 1 \sim \text{Cauchy}(\mu)$, $X|Y = 2 \sim \text{Cauchy}(-\mu)$. Now the distributions have heavy tails and $\eta(x) \to 1/2$ as $|x| \to \infty$.

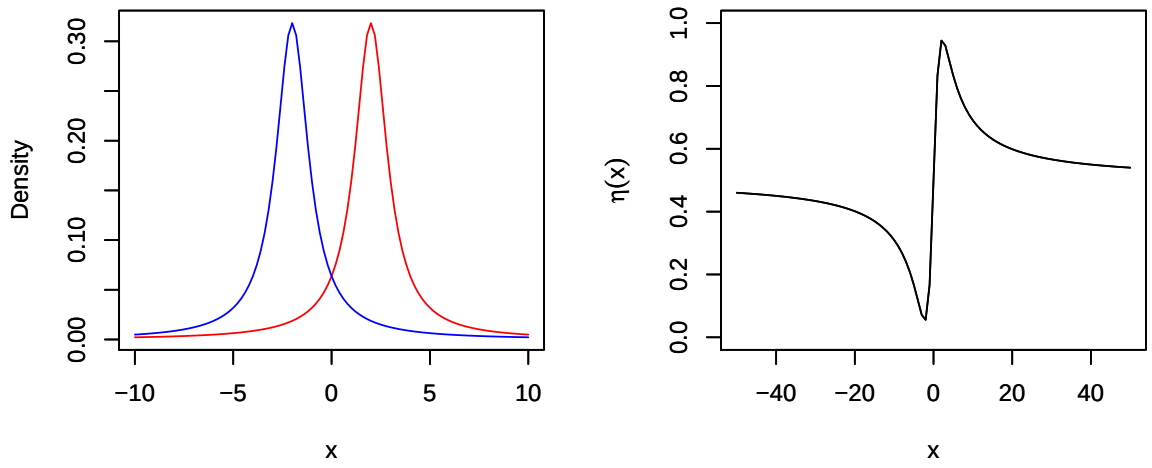

Figure 2: Left panel: The densities given $Y = 1$ (red) and $Y = 2$ (blue) for $\mu = 2$. Right panel: The corresponding regression function.

1.3.3 Cauchy Location Model 2

Suppose (X, Y) is such that $p = 3/4$ and $X|Y = 1 \sim \text{Cauchy}(\mu),\ X|Y = 2 \sim \text{Cauchy}(-\mu)$. In this case the prior probability is such that the regression function is bounded away from $1/2$ in the tail.

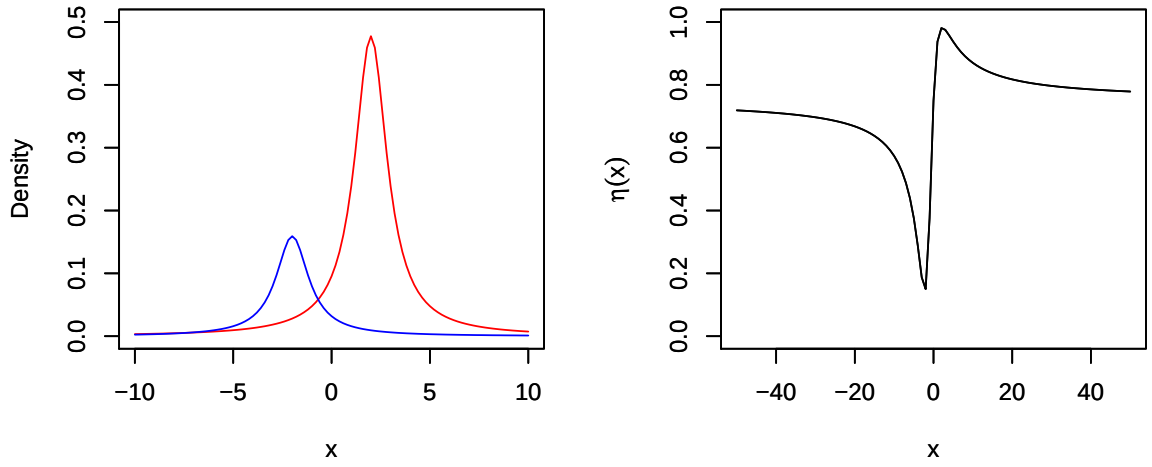

Figure 3: Left panel: The densities given $Y = 1$ (red) and $Y = 2$ (blue) for $\mu = 2$. Right panel: The corresponding regression function.

1.3.4 Sub-Cauchy Model

Suppose (X, Y) is such that $X|Y = 1 \sim f_1,\ X|Y = 2 \sim f_2$ where $f_1(x) = \frac{\sqrt{2}}{\pi\{1+(x-\mu)^4\}}$ and $f_2(x) = \frac{\sqrt{2}}{\pi\{1+(x+\mu)^4\}}$ with $p = 1/2$. Here $\eta(x) \to 1/2$ as $x \to \infty$ at the same rate as in the Cauchy example above but the tails are lighter.

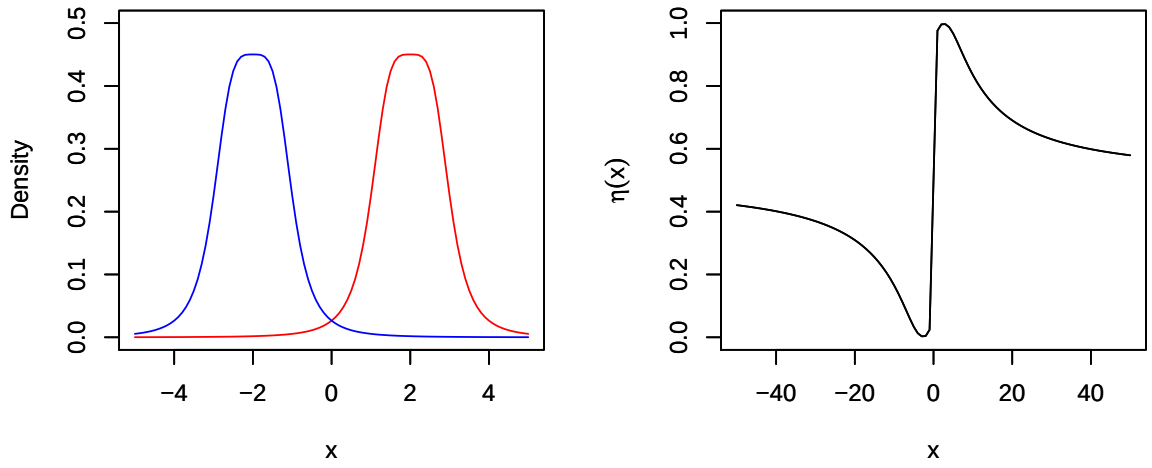

Figure 4: Left panel: The densities given $Y = 1$ (red) and $Y = 2$ (blue) for $\mu = 2$. Right panel: The corresponding regression function.

1.3.5 Pareto Model

Suppose (X, Y) is such that $X|Y = 1 \sim f_1,\ X|Y = 2 \sim f_2$ where $f_1(x) = (\alpha - 1)x^{-\alpha}\mathbb{1}_{\{x \geq 1\}}$, $f_2(x) = (\beta - 1)x^{-\beta}\mathbb{1}_{\{x \geq 1\}}$ and $p = 1/2$. In this example $\eta(x) \to 1$ as $x \to \infty$.

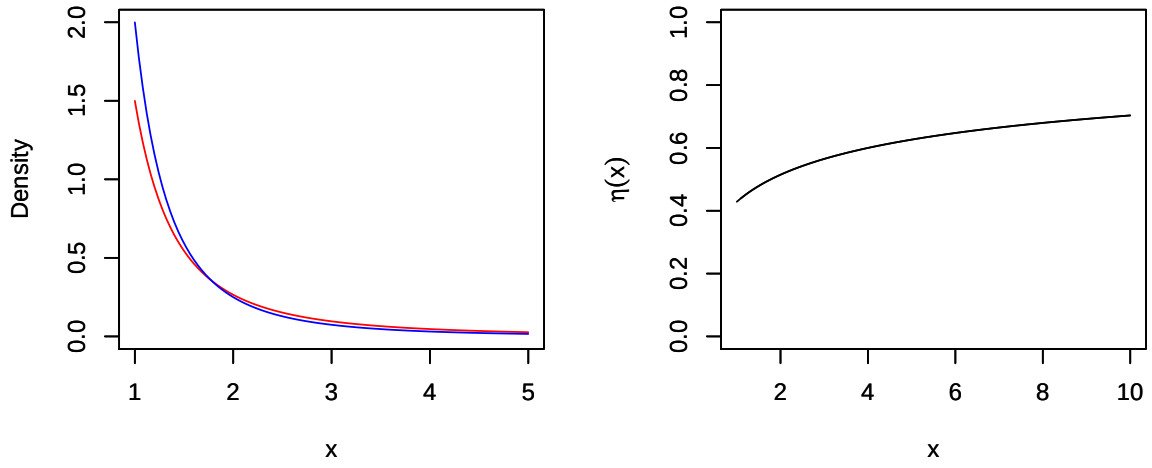

Figure 5: Left panel: The densities given $Y = 1$ ($\alpha = 1.5$) (red) and $Y = 2$ ($\beta = 2$) (blue). Right panel: The corresponding regression function.

1.3.6 Light Tailed Model

Suppose (X, Y) is such that $X|Y = 1 \sim \mathcal{N}(0, 1/2),\ X|Y = 2 \sim f_2$ where $f_2(x) = \frac{K_0}{1+\exp(x^2)}$, for some normalising constant K_0, and let p satisfy $\frac{p}{1-p} = \sqrt{\pi}K_0$ so that $\eta(x) \to 1/2$ as $x \to \pm\infty$. Notice $\eta(x) = \frac{1+\exp(x^2)}{1+2\exp(x^2)}$. In this final example we have light (normal type) tails, however the prior probabilities are such that $\eta(x) \to 1/2$ very fast in the tails.

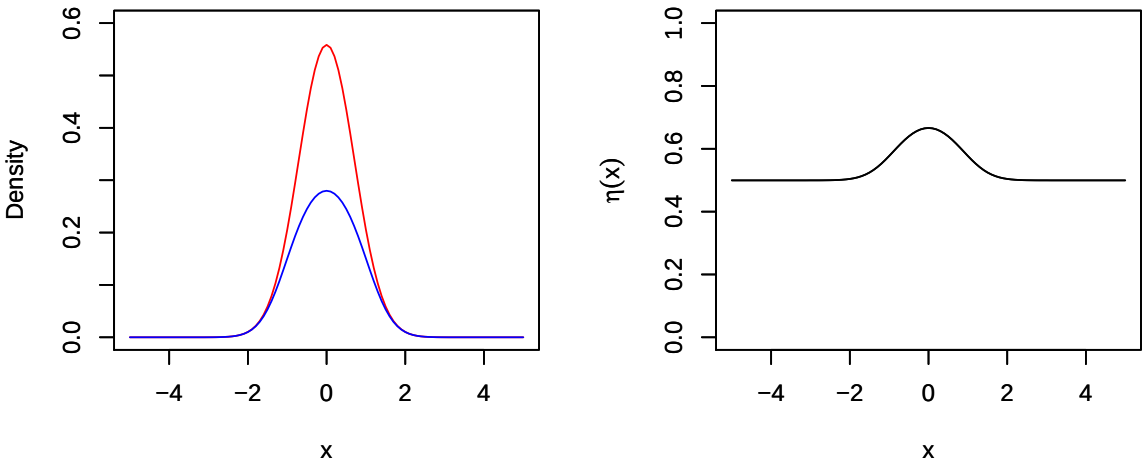

Figure 6: Left panel: The densities given $Y = 1$ (red) and $Y = 2$ (blue). Right panel: The corresponding regression function.

1.3.7 Discussion

Proposition 2.2 in Section 2 shows that when $\eta(x)$ is bounded away from $1/2$ as $x \to \infty$ (as in Examples 1.3.1, 1.3.3 and 1.3.5) then the contribution to the excess risk from the tails is of smaller order than that from the body of the distributions. If $\eta(x) \to 1/2$ as $x \to \infty$, there is a trichotomy of cases involving a trade off between the behaviour of $\eta(x)$ and the $\bar{P}$ measure in the tail.

- Case 1: The $\bar{P}$ measure of the tails is small in some sense compared to the rate that $\eta(x) \to 1/2$. Then the contribution will be of smaller order than that of the body of the distribution.
- Case 2: There is a balance between $\bar{P}$ and $\eta(x)$ such that we have an additional contribution of the same order.
- Case 3: The $\eta(x) \to 1/2$ fast in the tails, and $\bar{P}$ is not small enough to overcome this, then the dominant contribution the the excess risk is coming from the tails.

It is unclear into which case each of the earlier examples fits. In Theorems 2.3 and 2.4 in Section 2 we specify sufficient conditions to ensure we are in Case 1 (we will see that these conditions are satisfied in example 1.3.4). Furthermore, we show in Propositions 3.1 and 3.2 that the situation in example 1.3.2 may fit into Case 1 or 2 depending on the choice of the weight vector. Investigating Case 3 further is left to the subject of future work.

2 Classification in the Tails

In this section we attempt to relax assumption **(A. 1)**; we seek to obtain an asymptotic expansion for the excess risk over $\mathbb{R}^d$

$$\mathcal{R}_{\mathbb{R}^d}(\hat{C}_n^{wnn}) - \mathcal{R}_{\mathbb{R}^d}(C^{Bayes}),$$

as $n \to \infty$. There are very few results on this topic, indeed asymptotic expansions of the excess risk are usually restricted to a compact set (Mammen & Tsybakov, 1999; Audibert & Tsybakov, 2007; Biau et al., 2010; Samworth, 2012). Chanda & Ruymgaart (1989) and later Hall & Kang (2005) investigated the asymptotics in the tails for one-dimensional kernel-based classifiers. Hall & Kang (2005) use compactly supported kernels to classify univariate variables into one of two classes, where it is assumed that densities f_1 and f_2 exist. Their method first computes kernel density estimates $\hat{f}_1$ and $\hat{f}_2$ for f_1 and f_2 respectively, then classifies $x \in \mathbb{R}$ to group 1 if $p\hat{f}_1(x) > (1-p)\hat{f}_2(x)$ and group 2 otherwise. However both estimates $\hat{f}_1$ and $\hat{f}_2$ may be zero; this is common for x in the far tails of the conditional distributions. The authors suggest the following way of classifying such points: suppose x is in the right tail such that $\hat{f}_1(x) = \hat{f}_2(x) = 0$ and let

$$\hat{x} = \inf\{y : y \leq x \text{ and } \hat{f}_1(z) = \hat{f}_2(z) = 0 \text{ for all } z \in [y, x]\}.$$

Then, with probability one, exactly one of $\hat{f}_1(\hat{x}-)$ and $\hat{f}_2(\hat{x}-)$ will be non-zero - classify x to that class. It is shown that, under certain regularity conditions, this classification method performs well. The contribution from the tails to the excess risk is of smaller order than the contribution from the body of the distribution. The authors show further, however, that for Pareto type tails (specifically, $f_1(x) \sim ax^{-\alpha}$ and $f_2(x) \sim bx^{-\beta}$ as $x \to \infty$, where $a, b > 0$ and $1 < \alpha < \beta < \alpha + 1 < \infty$ as in example 1.3.5), the excess risk in the right tail is of larger order than that in the body of the distribution.

Nearest neighbour classifiers avoid the possibility of such ties in the tail, indeed in the extreme right tail the k nearest neighbour classifier will use the k largest order statistics from the training data to classify.

Recall the definition of the weighted nearest neighbour classifier in (1.3), and the regression estimate $S_n(x) = \sum_{i=1}^n w_{ni} \mathbb{1}_{\{Y_{(i)}=1\}}$. We will make use of the following equality: for any set $\mathcal{R} \subset \mathbb{R}^d$

$$\mathcal{R}_{\mathcal{R}}(\hat{C}_n^{wnn}) - \mathcal{R}_{\mathcal{R}}(C^{Bayes}) = \int_{\mathcal{R}} [\mathbb{P}\{S_n(x) < 1/2\} - \mathbb{1}_{\{\eta(x)<1/2\}}] P^{\circ}(dx). \tag{2.1}$$

2.1 One dimensional case

Here we discuss the univariate case where X takes values in $\mathbb{R}$. The following lemma gives bounds on the distance of the kth nearest neighbour to a point x in the tails.

Lemma 2.1. *Suppose x_0 is such that $p_{x_0/2}(x_0) := \bar{P}\{B_{x_0/2}(x_0)\} > 0$ and let $X_{(k)}(x)$ denote the kth nearest neighbour to x. Then for n sufficiently large,*

$$\mathbb{P}\{X_{(k)}(x) \notin [x_0/2, 2x] \text{ for some } x > x_0\} \leq \exp\left[-\frac{2\{np_{x_0/2}(x_0) - k\}^2}{n}\right]. \tag{2.2}$$

Proof. Observe for n large enough,

$$\begin{aligned}\mathbb{P}\{X_{(k)}(x) \notin [x_0/2, 2x] \text{ for some } x > x_0\ \} &\leq \mathbb{P}\{|X_{(k)}(x_0) - x_0| > x_0/2\} \\ &= \mathbb{P}[\text{Bin}\{n, p_{x_0/2}(x_0)\} < k] \\ &\leq \exp\left[-\frac{2\{np_{x_0/2}(x_0) - k\}^2}{n}\right].\end{aligned}$$

To see the first line note that if there are k points in $[\frac{1}{2}x_0, \frac{3}{2}x_0]$ then, for $x > x_0$, there are at least k in $[\frac{1}{2}x_0, 2x]$. Then we have used Hoeffding's inequality to bound the binomial probability. □

Lemma 2.1 gives us, with high probability, control of the maximum distance any point in the tails can be from its kth nearest neighbour. This control proves crucial to enable us to bound the excess risk in the far right tail. Consider now the case when $\eta(x)$ is bounded away from $1/2$ in the tail. Proposition 2.2 gives bounds on the excess risk in the tail under this condition.

Proposition 2.2. *Assume that there exists $\epsilon > 0$ and $x_0 \in \mathbb{R}$ such that for all $x > x_0/2$, $\eta(x) - \frac{1}{2} > \epsilon$. Then for each $\beta \in (0, 1/2)$ and each $M > 0$,*

$$\sup_{\mathbf{w}_n \in W_{n,\beta}} \{\mathcal{R}_{(x_0,\infty)}(\hat{C}_n^{wnn}) - \mathcal{R}_{(x_0,\infty)}(C^{Bayes})\} = \mathcal{O}(n^{-M}), \tag{2.3}$$

as $n \to \infty$.

Proof. Fix $M > 0$. If the support of $\bar{P}$ is compact there is nothing to prove. Therefore, without loss of generality, we may assume $p_{x_0/2}(x_0) > 0$. We seek to bound $\mathbb{P}\{S_n(x) < 1/2\}$. Letting

$\mu_n(x) = \mathbb{E}\{S_n(x)\}$, Lemma 2.1 yields the following:

$$\begin{aligned}
&\inf_{x>x_0} \mu_n(x) - 1/2 \\
&\quad \geq \inf_{x>x_0} \sum_{i=1}^{n} w_{ni} \mathbb{P}\{Y_{(i)} = 1 \cap X_{(k_2)} \in [x_0/2, 2x]\} - 1/2 \\
&\quad \geq \inf_{x>x_0} \left(\sum_{i=1}^{k_2} w_{ni} \right) (1/2 + \epsilon) \mathbb{P}\{X_{(k_2)} \in [x_0/2, 2x] \text{ for all } x > x_0\} - 1/2 \\
&\quad \geq \inf_{x>x_0} (1 - n^{-\beta/2})(1/2 + \epsilon) \left(1 - \exp\left[-\frac{2}{n} \{ n p_{x_0/2}(x_0) - k_2 \}^2 \right] \right) - 1/2 \\
&\quad \geq \epsilon/2,
\end{aligned}$$

for n large enough, uniformly for $\mathbf{w}_n \in W_{n,\beta}$. Here we have used the fact that $\inf_{u \in [x_0/2, 2x]} \eta(u) \geq 1/2 + \epsilon$. Now,

$$\begin{aligned}
\sup_{x \geq x_0} \mathbb{P}\{S_n(x) < 1/2\} &= \sup_{x \geq x_0} \mathbb{P}\{\mu_n(x) - S_n(x) > \mu_n(x) - 1/2\} \\
&\leq \sup_{x \geq x_0} \exp\left[-\frac{2}{s_n^2} \{\mu_n(x) - 1/2\}^2 \right] \\
&\leq \exp\left(-\frac{\epsilon^2}{2 s_n^2} \right) = \mathcal{O}(n^{-M}) \text{ for all } M > 0,
\end{aligned}$$

by Hoeffding's inequality, uniformly for $\mathbf{w}_n \in W_{n,\beta}$, since $s_n^2 \leq n^{-\beta}$ by assumption. It follows that

$$\begin{aligned}
&\sup_{\mathbf{w}_n \in W_{n,\beta}} \{\mathcal{R}_{(x_0,\infty)}(\hat{C}_n^{wnn}) - \mathcal{R}_{(x_0,\infty)}(C^{Bayes})\} \\
&\qquad = \sup_{\mathbf{w}_n \in W_{n,\beta}} \int_{x_0}^{\infty} \mathbb{P}\{S_n(x) < 1/2\} P^{\circ}(dx) \\
&\qquad = \mathcal{O}(n^{-M}).
\end{aligned}$$

□

Remarks

- Analogous results hold for $\eta(x) < 1/2$ and in the far left tail.
- This result does not require the tails of the distributions to be light, it only requires that $\eta(x)$ is eventually bounded away from a half. In the case where densities f_1 and f_2 exist for P_1 and P_2 respectively, we simply require the ratio of $pf_1(x)$ and $(1-p)f_2(x)$ to be bounded away from one in the tail.
- Note that given similar conditions in the left tail we may take our compact set in Theorem

1.3 to be $\mathcal{R} = [-x_0, x_0]$, and provided **(A.1)** - **(A. 4)** hold, the dominant contribution the the excess risk is coming from the Bayes decision boundary $\mathcal{S}$ since the contribution from the tail is $o(s_n^2 + t_n^2)$.

We might hope that even if the distributions are similar in the tails, so that $\eta(x) \to 1/2$ as $x \to \infty$, we will still be able to obtain *good* bounds on the contribution to the excess risk from this region. By imposing conditions on the how heavy the tails can be, we show that this is indeed the case. The following results give sufficient conditions for us to conclude that the contribution to the excess risk arising from the tails is $o(s_n^2)$. Firstly, Theorem 2.3 without the assumption that densities f_1 and f_2 exist for P_1 and P_2 in the tails, and secondly, Theorem 2.4 under similar, but more explicit, conditions in the case when densities do exist.

In Theorems 2.3 and 2.4 we make the assumption that there exists $x_0 \in \mathbb{R}$ such that $\eta(x) - 1/2 > 0$ for all $x > x_0$. When this is the case let η^+ and η^- be decreasing functions on $[x_0, \infty]$ such that $1/2 < \eta^-(x) \leq \eta(x) \leq \eta^+(x)$ for all $x > x_0$ and $\eta^+(x) \to 1/2$ as $x \to \infty$.

To ease notation we introduce the function

$$a(x) := \exp\left[-W\left\{\frac{2(\eta^-(2x) - 1/2)^2}{a_0}\right\}\right],$$

for some $a_0 > 1$, where $W : (-1/\mathrm{e}, 0) \to (-\infty, -1)$ returns the non-principal part of the so-called *Lambert-W* function satisfying $z = W(z)\exp\{W(z)\}$ as in Figure 7. Notice that $a(x) \to \infty$ as $x \to \infty$ since $W(z) \to -\infty$ as $z \to 0$ from below.

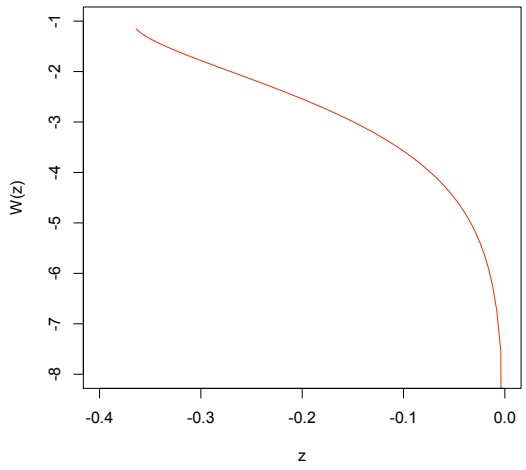

Figure 7: The non-principal part of the Lambert-W function for $z \in (-1/\mathrm{e}, 0)$.

Theorem 2.3. *Assume there exists x_0 such that $\eta(x) - \frac{1}{2} > 0$ for $x > x_0$ and that $\eta(x) \to \frac{1}{2}$ as*

$x \to \infty$. *Assume further that*

$$a(x)\{\eta^+(x) - 1/2\}\bar{P}([x,\infty)) \to 0,$$

as $x \to \infty$. *Then*

$$\mathcal{R}_{(x_0,\infty)}(\hat{C}_n^{wnn}) - \mathcal{R}_{(x_0,\infty)}(C^{Bayes}) = o(s_n^2),$$

as $n \to \infty$, *uniformly for* $\mathbf{w}_n \in W_{n,\beta}$.

Proof. Firstly, by conditioning on the training data $\mathcal{D}_n$, we have for any $x_1 \in \mathbb{R}$ that

$$\begin{aligned}\mathcal{R}_{(x_1,\infty)}(\hat{C}_n^{wnn}) - \mathcal{R}_{(x_1,\infty)}(C^{Bayes}) &= \mathbb{E}\left[2\int_{x_1}^{\infty}\{\eta(x) - 1/2\}\mathbb{1}_{\{S_n(x)<1/2\}}d\bar{P}(x)\middle|\mathcal{D}_n\right] \\ &\leq 2\int_{x_1}^{\infty}\{\eta(x) - 1/2\}d\bar{P}(x). \end{aligned} \tag{2.4}$$

Now choose an increasing sequence $x_1 := x_1(n)$ such that

$$\{\eta^-(2x_1) - 1/2\}^2 = -\left(\frac{a_0+1}{4}\right)s_n^2\log(s_n^2) =: \lambda_n,$$

for all $n > n_0$ where n_0 is such that $\sup_{x>x_0}\{\eta^-(2x) - 1/2\}^2 \geq -\left(\frac{a_0+1}{4}\right)s_{n_0}^2\log(s_{n_0}^2)$. From Lemma 2.1 we obtain,

$$\begin{aligned}&\inf_{x\in(x_0,x_1)}\mu_n(x) - \frac{1}{2} \geq \inf_{x\in(x_0,x_1)}\sum_{i=1}^{n} w_{n_i}\mathbb{P}\{Y_{(i)} = 1 \cap X_{(k_2)}(x) \in [x_0/2, 2x]\} - \frac{1}{2} \\ &\geq \inf_{x\in(x_0,x_1)}\left(\sum_{i=1}^{k_2} w_{n_i}\right)(1/2 + \lambda_n^{1/2})\mathbb{P}\{X_{(k_2)}(x) \in [x_0/2, 2x] \text{ for all } x > x_0\} - \frac{1}{2} \\ &\geq \inf_{x\in(x_0,x_1)}(1 - n^{-\beta})(1/2 + \lambda_n^{1/2})\left(1 - \exp\left[-\frac{2}{n}\{np_{x_0/2}(x_0) - k_2\}^2\right]\right) - \frac{1}{2} \\ &\geq \left[\left\{\frac{a_0+3}{2(a_0+1)}\right\}\lambda_n\right]^{1/2}, \end{aligned} \tag{2.5}$$

for n large enough, uniformly for $\mathbf{w}_n \in W_{n,\beta}$. Hence

$$\begin{aligned}\sup_{x\in(x_0,x_1)}\mathbb{P}\{S_n(x) < 1/2\} &= \sup_{x\in(x_0,x_1)}\mathbb{P}\{\mu_n(x) - S_n(x) > \mu_n(x) - 1/2\} \\ &\leq \sup_{x\in(x_0,x_1)}\exp\left[-\frac{2\{\mu_n(x) - 1/2\}^2}{s_n^2}\right] \\ &\leq \exp\left\{\left(\frac{a_0+3}{4}\right)\log(s_n^2)\right\} = o(s_n^2). \end{aligned} \tag{2.6}$$

as $n \to \infty$, uniformly for $\mathbf{w}_n \in W_{n,\beta}$. It follows that

$$\int_{x_0}^{x_1} \mathbb{P}\{S_n(x) < 1/2\} dP^\circ(x) = o(s_n^2), \tag{2.7}$$

as $n \to \infty$, uniformly for $\mathbf{w}_n \in W_{n,\beta}$. Further, for $n > n_0$

$$\begin{aligned}
&\frac{2}{s_n^2} \int_{x_1}^{\infty} \{\eta(x) - 1/2\} d\bar{P}(x) \\
&\qquad \leq \frac{2\{\eta^+(x_1) - 1/2\}}{s_n^2} \int_{x_1}^{\infty} d\bar{P}(x) \\
&\qquad \leq \exp\left(W\left[-\frac{2\{\eta^-(2x_1) - 1/2\}^2}{a_0} \right] \right) \{\eta^+(x_1) - 1/2\} \int_{x_1}^{\infty} d\bar{P}(x) \\
&\qquad = a(x_1)\{\eta^+(x_1) - 1/2\} \bar{P}([x_1, \infty)),
\end{aligned} \tag{2.8}$$

which converges to zero by assumption. The result follows from (2.7) and (2.8). □

If densities f_1 and f_2 exist for P_1 and P_2 let $\bar{f} = pf_1 + (1-p)f_2$, define

$$B(x) = \frac{\{\eta^+(x) - 1/2\}}{-\dot{\eta}^-(2x)\{\eta^-(2x) - 1/2\}} \bar{f}(x). \tag{2.9}$$

Theorem 2.4. *Assume that there exists x_0 such that $\eta(x) - 1/2 > 0$ for $x > x_0$ and $\eta(x) - 1/2 \to 0$ as $x \to \infty$. Assume further that for $x > x_0$ densities f_1 and f_2 exist for P_1 and P_2 respectively. Suppose further that $B(x) \to 0$ as $x \to \infty$. Then*

$$\mathcal{R}_{(x_0,\infty)}(\hat{C}_n^{wnn}) - \mathcal{R}_{(x_0,\infty)}(C^{Bayes}) = o(s_n^2), \tag{2.10}$$

as $n \to \infty$, uniformly for $\mathbf{w} \in W_{n,\beta}$.

Proof. Let $x_1 := x_1(n)$ be such that $\{\eta^-(2x_1) - 1/2\}^2 = -s_n^2 \log(s_n^2)$ (note that such a sequence exists for n sufficiently large). By a similar argument to that leading to (2.6) we have

$$\sup_{x \in (x_0, x_1)} \mathbb{P}\{S_n(x) < 1/2\} = o(s_n^2), \tag{2.11}$$

as $n \to \infty$, uniformly for $\mathbf{w}_n \in W_{n,\beta}$. To bound the contribution from the interval $[x_1, \infty]$, let $x_2 := x_2(n)$ be a sequence such that $\frac{\{\eta^-(2x_2) - 1/2\}^2}{s_n^2} = c$ for some constant $c > 0$ (note that for

all $c > 0$ and n large enough $x_2 \in (x_1, \infty)$). Then, by substituting $t = \frac{\{\eta^-(2x)-1/2\}^2}{s_n^2}$ we have

$$\begin{aligned}\sup_{\mathbf{w}_n \in W_{n,\beta}} \frac{2}{s_n^2} \int_{x_2}^{\infty} \{\eta(x) - 1/2\}\bar{f}(x)dx &\leq \sup_{\mathbf{w}_n \in W_{n,\beta}} \int_0^c B(x^t)/2dx \\ &\leq \sup_{\mathbf{w}_n \in W_{n,\beta}} \sup_{x \in (x_2,\infty)} \{B(x)\}c/2 \to 0. \end{aligned} \tag{2.12}$$

Furthermore, by a small modification to the argument leading to (2.5) we conclude

$$\sup_{\mathbf{w}_n \in W_{n,\beta}} \mu_n(x) - \frac{1}{2} \geq \frac{1}{2}\{\eta^-(2x) - 1/2\},$$

for n sufficiently large, uniformly for $x \in (x_1, x_2)$. Hence,

$$\mathbb{P}\{S_n(x) < 1/2\} \leq \exp\left[-\frac{\{\eta^-(2x) - 1/2\}^2}{2s_n^2}\right],$$

for n sufficiently large, uniformly for $\mathbf{w}_n \in W_{n,\beta}$ and $x \in (x_1, x_2)$. Finally, we make the substitution $t = \frac{\{\eta^-(2x)-1/2\}^2}{s_n^2}$ again to conclude that

$$\begin{aligned}&\sup_{\mathbf{w}_n \in W_{n,\beta}} \frac{1}{s_n^2} \int_{x_1}^{x_2} \mathbb{P}\{S_n(x) < 1/2\} dP^\circ(x) \\ &\quad\leq \sup_{\mathbf{w}_n \in W_{n,\beta}} \frac{2}{s_n^2} \int_{x_1}^{x_2} \exp\left[-\frac{\{\eta^-(2x) - 1/2\}^2}{2s_n^2}\right] \{\eta^+(x) - 1/2\}\bar{f}(x)dx \\ &\quad= \sup_{\mathbf{w}_n \in W_{n,\beta}} \int_{-\log(s_n^2)}^{c} \exp(-t/2) \left[\frac{\{\eta^+(x^t) - 1/2\}\bar{f}(x^t)}{\dot{\eta}^-(2x^t)\{\eta^-(2x^t) - 1/2\}}\right] dt \\ &\quad= \sup_{\mathbf{w}_n \in W_{n,\beta}} \int_{c}^{-\log(s_n^2)} \exp(-t/2)B(x^t)dt \\ &\quad\leq \sup_{\mathbf{w}_n \in W_{n,\beta}} \sup_{x \in (x_1,x_2)} \{B(x)\} \int_{c}^{-\log(s_n^2)} \exp(-t/2)dt \\ &\quad= \sup_{x \in (x_1,x_2)} \{B(x)\}2\{\exp(-c/2) - s_n\} \to 0 \end{aligned} \tag{2.13}$$

by assumption on $B(x)$ since $x_1 \to \infty$. The result follows from (2.11), (2.12) and (2.13). □

Corollary 2.5. *Assume the conditions of Theorem 2.4 except that $B(x) \to 0$. Suppose that η and $\bar{f}$ are decreasing for $x > x_0$ and the ratio $\frac{\bar{f}(x)}{-\dot{\eta}(x)} \to 0$. Then*

$$\mathcal{R}_{(x_0,\infty)}(\hat{C}_n^{wnn}) - \mathcal{R}_{(x_0,\infty)}(C^{Bayes}) = o(s_n^2), \tag{2.14}$$

as $n \to \infty$, uniformly for $\mathbf{w}_n \in W_{n,\beta}$.

Proof. Sketch; Since $\eta(x)$ and $\bar{f}(x)$ are decreasing we can bound $\mu_n(x) - 1/2$ below by $\eta(x) - 1/2$ uniformly for $x > x_0$. Hence we can replace $B(x)$ in (2.12) and (2.13) by $\frac{4\bar{f}(x)}{-\dot{\eta}(x)}$. □

The condition $B(x) \to 0$ as $x \to \infty$ amounts to the following: the faster $\eta(x) \to 1/2$, the lighter one requires the tails of $\bar{P}$. Note that the joint distribution of (X, Y) is specified by fixing $\bar{f}$ and $\eta(x)$. Moreover these choices can be made independently subject to $\eta(x) \in [0, 1]$ for all $x \in \mathbb{R}$, and $\bar{f}$ being a probability density function.

Example We return to the example in Section 1.3.4, $X|Y = 1 \sim f_1,\ X|Y = 2 \sim f_2$ with $f_1(x) = \frac{\sqrt{2}}{\pi\{1+(x-\mu)^4\}}$, $f_2(x) = \frac{\sqrt{2}}{\pi\{1+(x+\mu)^4\}}$ and $p = 1/2$. Here

$$B(x) \sim \frac{1/x}{(1/2x)^2(1/2x)} 1/x^4 \to 0, \text{ as } x \to \infty.$$

Hence for x_0 large enough we can apply Theorem 2.4 to conclude

$$\mathcal{R}_{(x_0,\infty)}(\hat{C}_n^{wnn}) - \mathcal{R}_{(x_0,\infty)}(C^{Bayes}) = o(s_n^2),$$

as $n \to \infty$, uniformly for $\mathbf{w}_n \in W_{n,\beta}$.

3 Cauchy Distributions

We investigate the results in the previous section further by appealing to a particular example, that is where the conditional distributions are location separated Cauchy distributions as in Example 1.3.2. We have $\mathbb{P}(Y = 1) = 1/2 = \mathbb{P}(Y = 2)$,

$$f_1(x) = \frac{1}{\pi\{1 + (x-\mu)^2\}},\ f_2(x) = \frac{1}{\pi\{1 + (x+\mu)^2\}}, \tag{3.1}$$

$$\bar{f}(x) = \frac{1 + x^2 + \mu^2}{\pi\{1 + (x-\mu)^2\}\{1 + (x+\mu)^2\}},$$

$$\eta(x) - 1/2 = \frac{\mu x}{1 + x^2 + \mu^2},$$

$$\dot{\eta}(x) = \frac{\mu(1 - x^2 + \mu^2)}{(1 + x^2 + \mu^2)^2}$$

and for $x > \mu$,

$$\bar{P}(x, \infty) = \frac{1}{2\pi}\left\{\tan^{-1}\left(\frac{1}{x-\mu}\right) + \tan^{-1}\left(\frac{1}{x+\mu}\right)\right\}.$$

Note that because η is decreasing in the right tail we can take $\eta^- = \eta = \eta^+$.

In this case $B(x) = \frac{\{\eta(x)-1/2\}}{-\dot{\eta}(2x)\{\eta(2x)-1/2\}}\bar{f}(x) \to \frac{16}{\pi\mu}$, so we cannot apply Theorem 2.4. However, by studying the proof of the theorem and the resulting corollary, we see that the contribution to the excess risk from the tail is bounded by

$$\frac{2c + 8\exp(-c/2)}{\pi\mu}s_n^2 = \frac{2 + 4\log 2}{\pi\mu}s_n^2,$$

for the optimal choice of c.

For simplicity we restrict use to unweighted k-nearest neighbour classifiers, i.e. $w_{ni} = \frac{1}{k}\mathbb{1}_{\{i\leq k\}}$. Fix $x_0 > 0$ and set $\mathcal{R} = [-x_0, x_0]$. From Theorem 1.3 we have, for each $\beta \in (0, 1/2)$,

$$\mathcal{R}_{\mathcal{R}}(\hat{C}_n^{knn}) - \mathcal{R}_{\mathcal{R}}(C^{Bayes}) = \frac{1}{4\pi\mu k}\{1 + o(1)\},$$

as $n \to \infty$ uniformly for $k \in [n^\beta, n^{1-\beta}]$. Note there is no contribution from the bias here since the problem is symmetric about zero. This poses the question:

Q: *Is the contribution to the excess risk arising from the tail, i.e. from outside any fixed compact set containing an open neighbourhood about zero, of the same order as that from a neighbourhood of the Bayes decision boundary $x = 0$ or can we obtain tighter bounds on the tail contribution and show it is $o(1/k)$?*

In the following two propositions we show that the answer depends on the choice of k. Define

$$\tilde{W}_{n,k} := \left\{\mathbf{w}_n : w_{ni} = \frac{1}{k}\mathbb{1}_{\{i\leq k\}}\right\}.$$

Proposition 3.1. *Suppose f_1 and f_2 are as in (3.1) and that $\mathbb{P}(Y = 1) = \mathbb{P}(Y = 2) = 1/2$. Suppose further $\mathbf{w}_n \in \tilde{W}_{n,k}$ and fix $x_0 > 0$. Then for each $\beta \in (0, 1/3)$*

$$\mathcal{R}_{(x_0,\infty)}(\hat{C}_n^{knn}) - \mathcal{R}_{(x_0,\infty)}(C^{Bayes}) = o(1/k),$$

as $n \to \infty$, uniformly for $k \in K_{1,\beta} := [n^{2/3}, n^{1-\beta}]$.

Proof. Let $x_3^- = \frac{n}{k\pi}\left(1 - \frac{1}{2\log n}\right)$ and $x_3^+ = \frac{n}{k\pi}\left(1 + \frac{1}{2\log n}\right)$. Then

$$\bar{P}(x_3^-, \infty) = \frac{1}{2\pi}\left\{\tan^{-1}\left(\frac{1}{x_3^- - \mu}\right) + \tan^{-1}\left(\frac{1}{x_3^- + \mu}\right)\right\} = \frac{1}{\pi x_3^-}\{1 + o(1)\},$$

and

$$\bar{P}(x_3^+, \infty) = \frac{1}{2\pi}\left\{\tan^{-1}\left(\frac{1}{x_3^+ - \mu}\right) + \tan^{-1}\left(\frac{1}{x_3^+ + \mu}\right)\right\} = \frac{1}{\pi x_3^+}\{1 + o(1)\}.$$

Suppose x is such that $\bar{P}(x,\infty) = o\left(\frac{k}{n\log n}\right)$. Arguing similarly the the proof of Lemma 2.1, it is straightforward to show that

$$\sup_{k\in K_{1,\beta}} \mathbb{P}\{X_{(k)}(x) < x_3^-\} = \mathcal{O}(n^{-M}),$$

and

$$\sup_{k\in K_{1,\beta}} \mathbb{P}\{X_{(k)}(x) > x_3^+\} = \mathcal{O}(n^{-M}).$$

Therefore, for such a choice of x, we have

$$\begin{aligned}
\mu_n(x) - 1/2 &= \frac{1}{k}\sum_{i=1}^{k} \mathbb{E}\{\eta(X_{(i)}) - 1/2\} \\
&= \mathbb{E}\left[\frac{1}{k}\sum_{i=1}^{n}\{\eta(X_{(i)}) - 1/2\}\mathbb{1}_{\{X_{(i)}\in(x_3^-,2x-x_3^-)\}}\right] + \mathcal{O}(n^{-M}) \\
&= \mathbb{E}\left[\frac{n}{k}\sum_{i=1}^{n}\frac{1}{n}\{\eta(X_i) - 1/2\}\mathbb{1}_{\{X_i\in(x_3^-,2x-x_3^-)\}}\right] + \mathcal{O}(n^{-M}) \\
&= \frac{n}{k}\int_{x_3^-}^{2x-x_3^-}\{\eta(t) - 1/2\}\bar{f}(t)dt\{1+o(1)\} \\
&= \frac{n}{k}\int_{x_3^-}^{\infty}\{\eta(t) - 1/2\}\bar{f}(t)dt\{1+o(1)\} \\
&= \frac{n}{2\pi k}\left\{\tan^{-1}\left(\frac{1}{x_3^- - \mu}\right) - \tan^{-1}\left(\frac{1}{x_3^- + \mu}\right)\right\}\{1+o(1)\} \\
&= \frac{n\mu}{2\pi k(x_3^{-2} - \mu^2)}\{1+o(1)\} \\
&= \frac{\pi\mu k}{n}\{1+o(1)\},
\end{aligned}$$

as $n\to\infty$, uniformly for $k\in K_{1,\beta}$. Hence by Hoeffding's inequality,

$$\mathbb{P}\{S_n(x) < 1/2\} \le \mathbb{P}\{\mu_n(x) - S_n(x) > \mu_n(x) - 1/2\} \le \exp\left[-k\{\mu_n(x) - 1/2\}^2\right] = o(1/k),$$

as $n\to\infty$, uniformly for $k\in K_{1,\beta}$.

Set $x_1 = \left(\frac{\mu k}{4\log k}\right)^{1/2}$. We have $\bar{P}(x_1,\infty) = o\left(\frac{k}{n\log n}\right)$ and therefore

$$\sup_{x\in(x_1,\infty)} \mathbb{P}\{S_n(x) < 1/2\} = o(1/k),$$

as $n \to \infty$, uniformly for $k \in K_{1,\beta}$. By a similar argument to that leading to (2.11)

$$\sup_{x \in (x_0, x_1)} \mathbb{P}\{S_n(x) < 1/2\} = o(1/k).$$

It follows that

$$\mathcal{R}_{(x_0,\infty)}(\hat{C}_n^{knn}) - \mathcal{R}_{(x_0,\infty)}(C^{Bayes}) = 2 \int_{x_0}^{\infty} \mathbb{P}\{S_n(x) < 1/2\}\{\eta(x) - 1/2\} \bar{f}(x) dx = o(1/k),$$

as $n \to \infty$, uniformly for $k \in K_{1,\beta}$. □

Proposition 3.2. *Suppose f_1 and f_2 are as in (3.1) and that $\mathbb{P}(Y = 1) = \mathbb{P}(Y = 2) = 1/2$. Suppose further $\mathbf{w}_n \in \tilde{W}_{n,k}$ and fix $x_0 > 0$. Then for each $\beta \in (1/3, 1/2)$*

$$\mathcal{R}_{(x_0,\infty)}(\hat{C}_n^{knn}) - \mathcal{R}_{(x_0,\infty)}(C^{Bayes}) = \frac{1}{2k\pi\mu}\{1 + o(1)\},$$

as $n \to \infty$, uniformly in $k \in K_{2,\beta} := [n^{\beta}, n^{1-\beta}]$.

Proof. Let $x_1 = \frac{k^{1/2}}{\log n}$ and $x_2 = k^{1/2} \log n$. For $x \in (x_1, x_2)$, set $x^+ = x\left(1 + \frac{1}{2\log n}\right)$ and $x^- = x\left(1 - \frac{1}{2\log n}\right)$. Then

$$\inf_{x \in (x_1, x_2)} \frac{n}{k \log n} \bar{P}(x^-, x^+) \to \infty,$$

and by similar arguments to that in Lemma 2.1, it can be shown that

$$\sup_{x \in (x_1, x_2)} \mathbb{P}\{X_{(k)}(x) \notin (x^-, x^+)\} = \mathcal{O}(n^{-M}).$$

Thus,

$$\mu_n(x) - 1/2 = \{\eta(x) - 1/2\}\{1 + o(1)\},$$

as $n \to \infty$, uniformly for $x \in (x_1, x_2)$ and $k \in K_{2,\beta}$.

By the Berry–Esseen theorem there exists C_0, such that for all $y \in \mathbb{R}$,

$$\sup_{k \in K_{2,\beta}} \sup_{x \in (x_1, x_2)} |\mathbb{P}[k^{1/2}\{S_n(x) - \mu_n(x)\} \le y] - \Phi(y)| \le \frac{C_0}{n^{1/2}(1 + |y|^3)},$$

where Φ denotes the standard normal distribution function. Hence

$$\mathbb{P}\{S_n(x) < 1/2\} = \Phi[-k^{1/2}\{\mu_n(x) - 1/2\}] + \mathcal{O}(n^{-1/2})$$

uniformly for $x \in (x_1, x_2)$ and $k \in K_{2,\beta}$. By making the substitution $t = k^{1/2}\{\eta(x) - 1/2\}$ we

have

$$
\begin{aligned}
2\int_{x_1}^{x_2} & \mathbb{P}\{S_n(x) < 1/2\}\{\eta(x) - 1/2\}\bar{f}(x)dx \\
&= 2\int_{x_1}^{x_2} \Phi[k^{1/2}\{1/2 - \mu_n(x)\}]\{\eta(x) - 1/2\}\bar{f}(x)dx\{1 + o(1)\} \\
&= 2\int_{x_1}^{x_2} \Phi[k^{1/2}\{1/2 - \eta(x)\}]\{\eta(x) - 1/2\}\bar{f}(x)dx\{1 + o(1)\} \\
&= \frac{2}{k}\int_0^{\infty} \frac{1}{\mu\pi} t\Phi\{-t\}dt\{1 + o(1)\} \\
&= \frac{1}{2k\pi\mu}\{1 + o(1)\},
\end{aligned}
$$

as $n \to \infty$, uniformly for $k \in K_{2,\beta}$.

It remains to bound the contribution to the excess risk from the intervals (x_0, x_1) and (x_2, ∞). Firstly, by a similar argument to that leading to (2.6)

$$\sup_{x \in (x_0, x_1)} \mathbb{P}\{S_n(x) < 1/2\} = o(1/k),$$

as $n \to \infty$ uniformly for $k \in K_{2,\beta}$. Secondly,

$$2\int_{x_2}^{\infty} \mathbb{P}\{S_n(x) < 1/2\}\{\eta(x) - 1/2\}\bar{f}(x)dx \le 2\int_{x_2}^{\infty} \{\eta(x) - 1/2\}\bar{f}(x)dx = o(1/k),$$

as $n \to \infty$, uniformly for $k \in K_{2,\beta}$. □

Remarks • Propositions 3.1 and 3.2 give an asymptotic expansion for the excess risk in the tail of the distribution for a k nearest neighbour classifier when the conditional distributions are Cauchy. We see that, depending on the choice on k, there are two cases. In the first case, for large k; that is when $\frac{k}{n^{2/3}} \to \infty$, it is shown that the excess risk is $o(1/k)$ and thus of smaller order than the that arising from and open neighbourhood of zero. For smaller k, when $\frac{k}{n^{2/3}}$ is bounded, the asymptotic expansion in the right tail is $\frac{\mu}{2\pi k}\{1 + o(1)\}$. The dominant term in this expansion is of the same order as the expansion over a neighbourhood of zero. For example when $k = n^{4/5}$ (the *optimal* choice for classification over a compact set), we have shown that

$$\mathcal{R}_{(x_0,\infty)}(\hat{C}_n^{knn}) - \mathcal{R}_{(x_0,\infty)}(C^{Bayes}) = o(n^{-4/5}),$$

as $n \to \infty$. However if $k = n^{1/2}$ say, then

$$\mathcal{R}_{(x_0,\infty)}(\hat{C}_n^{knn}) - \mathcal{R}_{(x_0,\infty)}(C^{Bayes}) = \frac{1}{2\mu\pi n^{1/2}}\{1 + o(1)\},$$

as $n \to \infty$.

- Due to the symmetry of the problem in this case we have the same conclusions for the region $(-\infty, -x_0)$.

4 Higher Dimensions

The situation for $d \geq 2$ is somewhat different, here we may have situations where the Bayes decision boundary $\mathcal{S} = \{x : \eta(x) = \frac{1}{2}\}$ extends to infinity (i.e. for every compact set $\mathcal{R}$ in $\mathbb{R}^d$ there exists $x \notin \mathcal{R}$ such that $\eta(x) = 1/2$).

To investigate this problem further we discuss the special case when $X|Y = 1 \sim \mathcal{N}_d(\boldsymbol{\mu}, I)$ and $X|Y = 2 \sim \mathcal{N}_d(-\boldsymbol{\mu}, I)$ with $\boldsymbol{\mu} = (\mu, 0, \ldots, 0)$, here $\eta(x) = \frac{1}{1+\exp(-\mu x_1)}$.

Proposition 4.1. *Suppose $d \geq 5$ and assume the normal location setting above. Suppose also we use a weighted nearest neighbour classifier with optimal weights $\mathbf{w}_n^*$ given by (1.6). Then*

$$\mathcal{R}_{\mathbb{R}^d}(\hat{C}_n^{wnn}) - \mathcal{R}_{\mathbb{R}^d}(C^{Bayes}) = \gamma_n(\mathbf{w}_n^*)\{1 + o(1)\} = \mathcal{O}\left(n^{-\frac{4}{d+4}}\right),$$

as $n \to \infty$.

Proof. Sketch; The proof here follows from a slight modification to the proof in Samworth (2012). The optimal weights $\mathbf{w}_n^*$ are contained in W_{n,β_d}, where $\beta_d = \frac{4}{d+4}$. Let $\mathcal{R}_n = \{x : \|x\|^2 \leq 2\left(\frac{4}{d+4} + \epsilon\right)\log n\}$ for some $\epsilon > 0$ small. Then by a similar argument to that in the proof of Theorem 1 in Samworth (2012), we can ensure that, uniformly for $x \in \mathcal{R}_n$, $X_{(k_2)}$ is contained in a small ball centred at x. We conclude that $X_{(k_2)}$ is the *correct* side of the boundary with high probability. Further, it can be shown that

$$\mathcal{R}_{\mathcal{R}_n^c}(\hat{C}_n^{wnn}) - \mathcal{R}_{\mathcal{R}_n^c}(C^{Bayes}) \leq \mathbb{P}\left\{\chi_d^2 \geq 2\left(\frac{4}{d+4} + \epsilon\right)\log n\right\} = o\left(n^{-\frac{4}{d+4}}\right).$$

□

Remarks

- At present we have no results for $d = 2, 3$ or 4. Proposition 4.1 falls down in these cases since we require the sequence of compact sets $\mathcal{R}_n$ to be growing at a faster rate and can no longer guarantee that uniformly for $x \in \mathcal{R}_n$, the k_2th nearest neighbour is close to x.

5 Discussion

We have investigated the properties of weighted nearest neighbour classifiers where we are interested in classifying points from the whole of $\mathbb{R}^d$. We showed in Section 2.1 that, under certain regularity conditions, classification in the tails of the marginal distribution of a univariate variable X is 'easy'. From the results in this section we can conclude that the asymptotic expansion of the excess risk derived by Samworth (2012) remains valid over $\mathbb{R}$, moreover, the weights specified in (1.6) are asymptotically optimal for the task presented.

Without such regularity conditions we have shown that there are situations where the dominant term in the asymptotic expansion of the excess risk from the tail of the distribution is of the same order as that from the body. In this case the optimal weighting scheme will be modified slightly; specifically the constant in the definition of k^* in (1.6) will be different.

This paper is left with the following open questions:

- For $d = 1$, can we find an explicit expression for the additional contribution to the excess risk for a general joint distribution for (X, Y);
- and determine exactly when the dominant contribution to the excess risk is arising from the tails, and specify optimal weights in this case.
- What can we say if η is oscillating about $1/2$ in the tail?
- For $d \geq 2$, determine conditions under which we can we can derive analogous results to those in Section 2.1.

References

Audibert, J.-Y., & Tsybakov, A. B. (2007). Fast learning rates for plug-in classifiers. *Ann. Statist.*, *35*, 608–633.

Biau, G., Cérou, F., & 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. (2010). On the layered nearest neighbour estimate, the bagged nearest neighbour estimate and the random forest method in regression and classification. *J. Mult. Anal.*, *101*, 2499–2518.

Chanda, K. C., & Ruymgaart, F. (1989). Asymptotic estimate of probability of misclassification for discriminant rules based on density estimates. *Statist Probab. Lett.*, *8*, 81 – 88.

Cover, T. M., & Hart, P. E. (1967). Nearest neighbour pattern classification. *IEEE Trans. Inf. Th.*, *13*, 21–27.

Devroye, L., Gyorfi, L., & Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*. Springer.

Fix, E., & Hodges, J. L. (1951). *Discriminatory analysis – nonparametric discrimination: Consistency properties*. Technical Report 4 USAF School of Aviation Medicine Randolph Field, Texas.

Hall, P., & Kang, K.-H. (2005). Bandwidth choice for nonparametric classification. *Ann. Statist.*, *33*, 284–306.

Hall, P., & 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.

Mammen, E., & Tsybakov, A. B. (1999). Smooth discriminant analysis. *Ann. Statist.*, *27*, 1808–1829.

Marron, J. S. (1983). Optimal rates of convergence to Bayes risk in nonparametric discrimination. *Ann. Statist.*, *11*, 1142–1155.

Samworth, R. J. (2012). Optimal weighted nearest neighbour classifiers. *Ann. Statist.*, *To appear*.

Stone, C. J. (1977). Consistent nonparametrc regression. *Ann. Statist.*, *5*, 595–620.

Tsybakov, A. B. (2005). Optimal aggregation of classifiers in statistical learning. *Ann. Statist.*, *32*, 135–166.