Journal of Machine Learning Research 26 (2025) 1-46 Submitted 10/23; Revised 7/25; Published 9/25

A Hybrid Weighted Nearest Neighbour Classifier for Semi-Supervised Learning

Stephen M. S. Lee SMSLEE@HKU.HK
Department of Statistics and Actuarial Science
The University of Hong Kong
Pokfulam Road, Hong Kong

Mehdi Soleymani* MEHDI.SOLEYMANI@ME.COM
On Lake
44388 Dortmund
Germany

Editor: Nicolas Vayatis

Abstract

We propose a novel hybrid procedure for constructing a randomly weighted nearest neighbour classifier for semi-supervised learning. The procedure first uses the labelled learning set to predict a probability distribution of class labels for the unlabelled learning set. This turns the unlabelled set into a pseudo-labelled set, on which a sequentially weighted nearest neighbour classifier can be trained. The vote proportions calculated by this sequentially weighted nearest neighbour classifier and the standard weighted nearest neighbour classifier trained on the labelled set alone are then linearly combined to build a hybrid classifier. Our theory shows that, given a sufficiently large set of unlabelled data, the hybrid classifier has an optimal regret converging at a faster rate than that of the optimally weighted nearest neighbour classifier and hence of the optimal bagged or k-nearest neighbour classifier. We also show that the hybrid classifier can be revised by a dislabelling strategy to achieve the fastest possible rate of regret irrespective of the size of the unlabelled set, which may even be empty. Simulation studies and real data examples are presented to support our theoretical findings and illustrate the empirical performance of the hybrid classifiers constructed using uniform weights. We also explore the effects of pseudo-labelling by hypothesized class probabilities as a supplement to our main findings.

Keywords: classification, nearest neighbour, semi-supervised learning, machine learning, optimal rate.

1. Introduction

Supervised classification aims to predict the class of a test point $x \in \mathcal{X}$, based on algorithms trained on a learning set consisting of n data points, $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \in \mathcal{X}$, which have been categorised into known classes. The learning set is said to be labelled. A popular algorithm for supervised classification is provided by the k-nearest neighbour classifier, introduced by Fix and Hodges (1951), which assigns a test point x to the jth class if a majority of the k learning observations nearest to x belong to the jth class. It has been applied to a wide range of statistical problems including, for example, gene classification (Li et al., 2001) and object recognition (Belongie et al., 2002). More generally, the k-nearest neighbour classifier

can be viewed as a special case of a weighted nearest neighbour classifier, on which we shall focus in this paper. The weighted nearest neighbour classifier places heavier weights on learning observations which are closer to x. Dudani (1976) demonstrates the benefits given by placing non-uniform, distance-dependent, weights on the k nearest neighbours. Under some mild assumptions, the weighted nearest neighbour classifier is Bayes consistent (Stone, 1977; Gadat et al., 2016). The choice of weights has been the focus of many works. Biau and Devroye (2015) survey the key ideas and state the consistency theorems for weighted nearest neighbour classifiers. Samworth (2012) derives a formula for the optimal weight vector which minimises asymptotically the misclassification rate in excess of the Bayes risk, among all choices of deterministic weight vectors. Berrett and Samworth (2019) propose an efficient two-sample functional estimator based on weighted nearest neighbours.

Recent years have seen increasing attention paid to semi-supervised classification, which finds applications in many machine learning problems where a set of unlabelled data points is available in addition to a labelled learning set, with both the labelled and unlabelled sets drawn from a common marginal distribution. Generally speaking, semi-supervised learning seeks to leverage information extracted from the unlabelled data to improve upon supervised learning based solely on the labelled data. Development of semi-supervised methods has become all the more imperative when many contemporary applications have found an abundant supply of unlabelled data, while labelled data are difficult or expensive to acquire. A general introduction to semi-supervised learning and a survey on its advances can be found in Zhu and Goldberg (2009) and van Engelen and Hoos (2019), respectively. In the context of classification, a vast amount of works exist across different fields regarding the application of semi-supervised learning. One common approach makes use of the unlabelled data to modify and strategically optimise a loss function: see, for example, Vapnik (1998), Joachims (1999), Berthelot et al. (2019) and Rebuffi et al. (2020). Other approaches consider semi-supervised classification under specialised settings. Examples include co-training (Blum and Mitchell, 1998) and parametric approaches such as EM-based incomplete likelihood maximisation under generative models (Nigam et al., 2000) or minimum entropy regularisation under discriminative models (Grandvalet and Bengio, 2004).

Given its long-standing prominence in the standard methodology of nonparametric supervised learning, nearest neighbour classification has naturally aroused research interests in its potential extensions to a nonparametric semi-supervised setting. Such extensions have been studied by, for example, Wang et al. (2010), Wajeed and Adilakshmi (2011), Liu et al. (2013) and Tu et al. (2016). None of the above works, however, consider formal statistical properties of their proposed methods. One exception is Cannings et al. (2020), who estimate the marginal density of the feature vector based on an unlabelled learning set and use the estimate to construct a local choice of k for a k-nearest neighbour algorithm. They show that their classifier has an excess risk converging at the minimax rate under weaker conditions, provided that the size of the unlabelled set diverges at some rate faster than n^2.

Soleymani and Lee (2014) propose a sequential algorithm for bagging a generic supervised classifier. When applied to the nearest neighbour method, their algorithm gives rise to a randomly weighted nearest neighbour classifier which is more stable than the conventional bagged nearest neighbour classifier. Inspired by their findings, we propose in this paper a hybrid weighted nearest neighbour procedure for semi-supervised classification, and

show that the classifier, if optimally tuned, yields a faster convergence rate for the excess risk compared to that of any deterministically weighted nearest neighbour classifier. To construct the hybrid classifier, we first train a weighted nearest neighbour classifier on the labelled data and use it to assign pseudo-labels, in the form of class probabilities, to the unlabelled data. Next we build a sequentially weighted nearest neighbour classifier in the manner of Soleymani and Lee (2014), by training another nearest neighbour classifier on the pseudo-labelled data. Our hybrid classifier is finally constructed by linearly combining the nearest neighbour classifier trained on the labelled data and the sequentially weighted nearest neighbour classifier trained on the pseudo-labelled data. Our theory shows that the performance of the hybrid classifier is determined critically by the weights employed in the two constituent weighted nearest neighbour classifiers, as well as the coefficients used for combining the two classifiers. It also enables us to derive the optimal tuning parameters which minimise the misclassification rate of the hybrid classifier.

The rest of the paper is organised as follows. Section 2 introduces our proposed hybrid weighted nearest neighbour classifier. Section 3 illustrates with a toy example the empirical performance of the hybrid classifier in comparison with its two constituents. Section 4 establishes theoretical results on the regret, or excess risk, of our hybrid classifier when its constituent weighted nearest neighbour classifiers are constructed using exponential and uniform weights, respectively. Section 5 presents empirical results obtained from simulation studies and three real data examples. Section 6 concludes our findings. All proofs are given in the Appendix.

2. Hybrid weighted nearest neighbour classifier

Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n, \boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_m$ denote independent data points drawn from K possible distributions on $\mathcal{X} \subset \mathbb{R}^d$. To each $\boldsymbol{X}_i$ is attached a class label $Y_i \in \{1, \ldots, K\}$ if $\boldsymbol{X}_i$ is drawn from the Y_i-th distribution. Let $\mathcal{L} = \mathcal{L}_S \cup \mathcal{L}_U$ be a semi-supervised learning set, where $\mathcal{L}_S = \{(\boldsymbol{X}_i, Y_i) : i = 1, \ldots, n\}$ and $\mathcal{L}_U = \{\boldsymbol{Z}_i : i = 1, \ldots, m\}$ denote the labelled and unlabelled parts of the learning set, respectively.

Consider for simplicity a two-class problem with $K = 2$. For any $x \in \mathcal{X}$, let $\big\{(\boldsymbol{X}_{(i)}(x), Y_{(i)}(x)) : i = 1, \ldots, n\big\}$ be a permutation of $\mathcal{L}_S$ such that $\boldsymbol{X}_{(i)}(x)$ is the ith nearest point among $\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}$ to x. Similarly, $\boldsymbol{Z}_{(i)}(x)$ denotes the ith nearest point among $\mathcal{L}_U$ to x. For a given vector of weights $\boldsymbol{w}_n = (w_{n,1}, \ldots, w_{n,n})$ with $w_{n,i} \geq 0$ and $\sum_{i=1}^n w_{n,i} = 1$, a weighted nearest neighbour classifier $T_{\boldsymbol{w}_n}$, trained on $\mathcal{L}_S$, assigns x to the first population if the weighted vote proportion $\hat{p}(x) = \sum_{i=1}^n w_{n,i} \mathbf{1}\{Y_{(i)}(x) = 1\} > 1/2$, and to the second population otherwise, where $\mathbf{1}\{\cdot\}$ denotes the indicator function. Formally, we define $T_{\boldsymbol{w}_n}(\mathcal{L}_S, x) = 1 + \mathbf{1}\big\{\hat{p}(x) \leq 1/2\big\}$. Using the above formulation, the k-nearest neighbour classifier corresponds to the choice of uniform weights $w_{n,i} = k^{-1}\mathbf{1}\{i \leq k\}$. Stone (1977) obtains sufficient conditions on $\boldsymbol{w}_n$ for L^r-convergence of $\hat{p}(x)$ to the conditional class probability $\mathbb{P}(Y_1 = 1 | \boldsymbol{X}_1 = x)$. The M out of n bagged nearest neighbour classifiers, constructed using infinitely many bootstrap resamples drawn from $\mathcal{L}_S$, correspond to setting

$$w_{n,i} = \Big(1 - \frac{i-1}{n}\Big)^M - \Big(1 - \frac{i}{n}\Big)^M, \;\; i = 1, \ldots, n, \tag{1}$$

if resampling is done with replacement, or

$$w_{n,i} = \binom{n-i}{M-1}\binom{n}{M}^{-1} \mathbf{1}\{i \leq n-M+1\}, \quad i = 1, \ldots, n, \tag{2}$$

if resampling is done without replacement. Hall and Samworth (2005) show that the M out of n bagged nearest neighbour classifier converges to the Bayes rule provided that $M \to \infty$ and $M/n \to 0$. Biau et al. (2010) show that the optimal rate of convergence of the vote proportions can be achieved by the k-nearest neighbour classifier or the M out of n bagged nearest neighbour classifier if k or M is chosen to diverge at an appropriate rate.

Given any weight vectors $\boldsymbol{w}_n = (w_{n,1}, \ldots, w_{n,n})$ and $\boldsymbol{w}_m^* = (w_{m,1}^*, \ldots, w_{m,m}^*)$, we train a sequentially weighted nearest neighbour classifier (Soleymani and Lee, 2014) on the semi-supervised learning set $\mathcal{L}$ as follows. By training $T_{\boldsymbol{w}_n}$ first on the labelled set $\mathcal{L}_S$, we calculate for each point $\boldsymbol{Z}_j$ in the unlabelled set $\mathcal{L}_U$ the weighted vote proportion $\hat{p}(\boldsymbol{Z}_j)$, for $j = 1, \ldots, m$. Treating the $\hat{p}(\boldsymbol{Z}_j)$'s as pseudo-labels for $\mathcal{L}_U$, the weighted nearest neighbour classifier $T_{\boldsymbol{w}_m^*}$ is trained next on $\mathcal{L}_U$ to predict the class of any given $x \in \mathcal{X}$. Thus, the sequentially weighted nearest neighbour classifier assigns x to the first population if and only if the sequentially weighted vote proportion $\hat{s}(x) = \sum_{j=1}^m w_{m,j}^* \hat{p}(\boldsymbol{Z}_{(j)}(x)) > 1/2$. The crux of the above sequential construction is the labelling of the data points in the unlabelled set $\mathcal{L}_U$ by class probabilities $\hat{p}(\boldsymbol{Z}_j)$ estimated using the labelled set $\mathcal{L}_S$, which has the effect of stabilising the weighted nearest neighbour classifier.

Define, for a tuning parameter $\varsigma \in \mathbb{R}$, a hybrid weighted nearest neighbour classifier to be

$$\mathscr{T}_{n,m,\varsigma}(\mathcal{L}, x) = 1 + \mathbf{1}\big\{\varsigma\hat{p}(x) + (1-\varsigma)\hat{s}(x) \leq 1/2\big\},$$

which assigns x to class 1 if and only if a linear combination of the two weighted vote proportions of class 1, calculated respectively by $T_{\boldsymbol{w}_n}$ and the sequentially weighted nearest neighbour classifier, exceeds 1/2. Note that the above linear combination is not restricted to be convex. In fact, we shall show later that the tuning parameter ς has an optimal value bigger than 1.

We may view $\mathscr{T}_{n,m,\varsigma}$ as a generalised version of a weighted nearest neighbour classifier $T_{\boldsymbol{w}_n^{**}}(\mathcal{L}_S, \cdot)$, with weights $\boldsymbol{w}_n^{**} = (w_{n,1}^{**}, \ldots, w_{n,n}^{**})$ given by

$$w_{n,i}^{**} = \varsigma w_{n,i} + (1-\varsigma)\sum_{i'=1}^{n}\sum_{j=1}^{m} w_{n,i'} w_{m,j}^* \, \mathbf{1}\big\{\boldsymbol{X}_{(i')}(\boldsymbol{Z}_{(j)}(x)) = \boldsymbol{X}_{(i)}(x)\big\}, \quad i = 1, \ldots, n,$$

which are random, depend on $(x, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n, \boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_m)$, satisfy $\sum_{i=1}^n w_{n,i}^{**} = 1$ and may assume negative values.

Figure 1 illustrates the workflow of our hybrid classifier $\mathscr{T}_{n,m,\varsigma}$, with $T_{\boldsymbol{w}_n}$ and $T_{\boldsymbol{w}_m^*}$ set to be the 3- and 4-nearest neighbour classifiers, respectively. At the pseudo-labelling step, the four unlabelled points nearest to the test point x are assigned by $T_{\boldsymbol{w}_n}$ the vote proportions $\hat{p}\big(\boldsymbol{Z}_{(1)}(x)\big) = 1/3$ and $\hat{p}\big(\boldsymbol{Z}_{(2)}(x)\big) = \hat{p}\big(\boldsymbol{Z}_{(3)}(x)\big) = \hat{p}\big(\boldsymbol{Z}_{(4)}(x)\big) = 2/3$. This gives rise to the sequentially weighted vote proportion $\hat{s}(x) = 4^{-1}\sum_{j=1}^4 \hat{p}\big(\boldsymbol{Z}_{(j)}(x)\big) = 7/12$ for x. Combining $\hat{s}(x)$ with the vote proportion $\hat{p}(x) = 1/3$ predicted by $T_{\boldsymbol{w}_n}$, our hybrid classifier $\mathscr{T}_{n,m,\varsigma}$ would assign x to class 1 if $\varsigma/3 + 7(1-\varsigma)/12 > 1/2$.

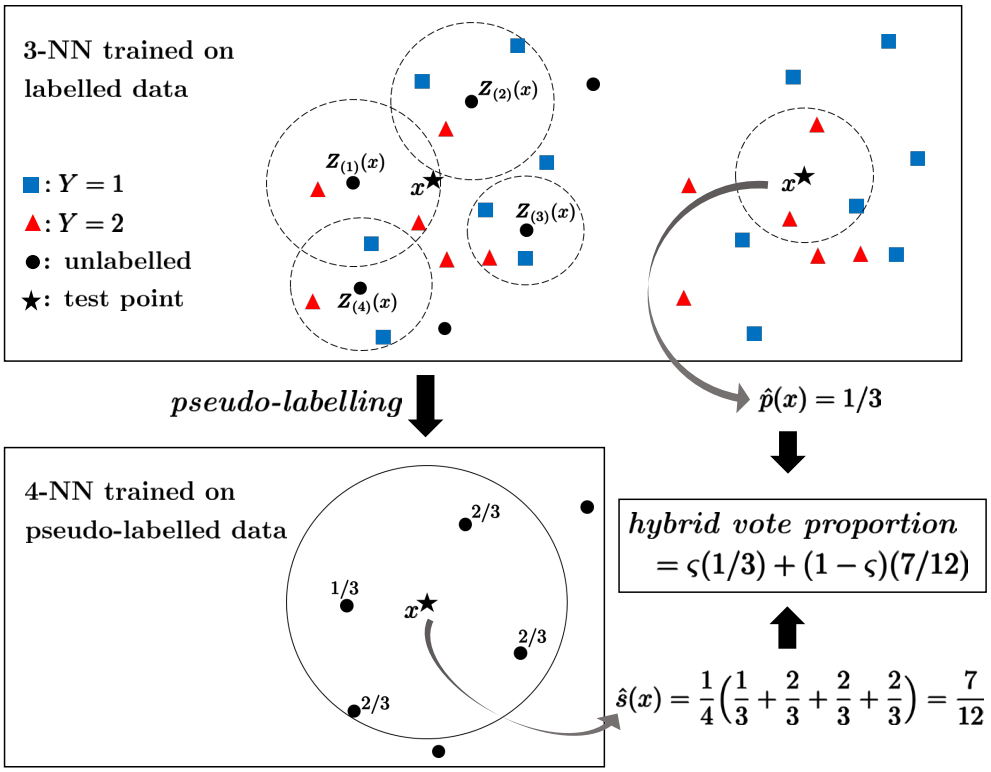

Figure 1: Workflow of hybrid weighted nearest neighbour classification of test point x, based on 3- and 4-nearest neighbour classifiers.

3. A toy experiment

We conduct a numerical experiment to compare the hybrid classifier $\mathscr{T}_{n,m,\varsigma}$ with its two constituents, that is the weighted nearest neighbour classifier $T_{\boldsymbol{w}_n}$ and the sequentially weighted nearest neighbour classifier derived from $T_{\boldsymbol{w}^*_m}$, in an application to the classical two-moon toy data set. We set $w_{n,i} = k_1^{-1}\mathbf{1}\{i \le k_1\}$ and $w^*_{m,i} = k_2^{-1}\mathbf{1}\{i \le k_2\}$, so that $T_{\boldsymbol{w}_n}$ and $T_{\boldsymbol{w}^*_m}$ amount to the k_1- and k_2-nearest neighbour classifiers, respectively. Recall that a test point x is assigned class 1 by $T_{\boldsymbol{w}_n}$ if $\hat{p}(x) > 1/2$, by the sequentially weighted nearest neighbour classifier if $\hat{s}(x) > 1/2$ and by $\mathscr{T}_{n,m,\varsigma}$ if $\varsigma\hat{p}(x) + (1-\varsigma)\hat{s}(x) > 1/2$. The tuning parameters (k_1, k_2) are selected from a grid of pilot values, which may differ between different methods. The mixing coefficient ς is set to be $1 + (k_1 m)/(k_2 n)$, an optimal choice for the two-dimensional case as derived in Section 4.3.

The two-moon labelled data are generated by setting $\mathbb{P}(Y = 1) = \mathbb{P}(Y = 2) = 1/2$ and

$$\boldsymbol{X} = \mathbf{1}\{Y = 1\}\begin{bmatrix}\cos U \\ \sin U\end{bmatrix} + \mathbf{1}\{Y = 2\}\begin{bmatrix}1 - \cos U \\ 0.5 - \sin U\end{bmatrix} + 0.5\,\boldsymbol{\epsilon},$$

for U uniformly distributed over $[0, \pi]$ and $\boldsymbol{\epsilon}$ bivariate standard normal, with $Y, U, \boldsymbol{\epsilon}$ independent of each other.

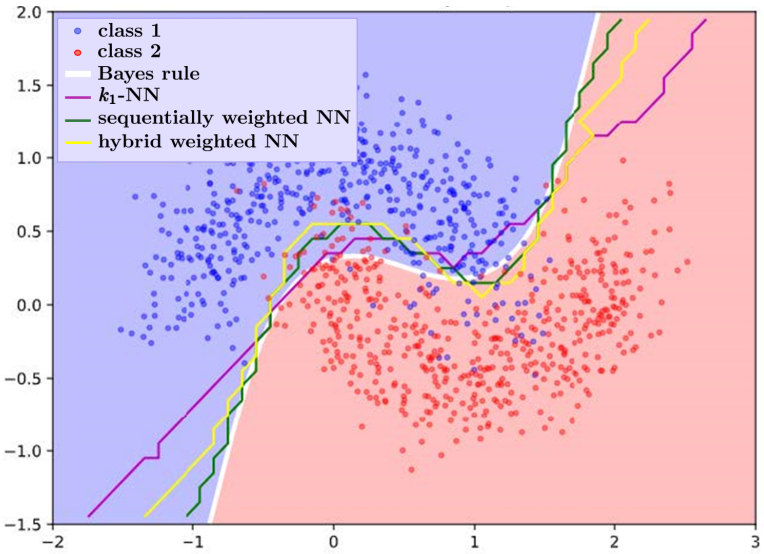

Figure 2: Two-moon data example — classification boundaries of Bayes rule, k_1-nearest neighbour classifier, sequentially weighted nearest neighbour classifier and hybrid weighted nearest neighbour classifier, estimated by averaging over 50 simulated semi-supervised learning sets, each consisting of 10 labelled and 100 unlabelled data points. A typical set of labelled data is shown in the background for reference.

Figure 2 displays a typical set of labelled data, with the blue and red points generated from the first and second populations, respectively. The white curve on the figure indicates the classification boundary of the Bayes rule, given by the set $\{x : \mathbb{P}(Y = 1|\boldsymbol{X} = x) = 1/2\}$. The figure also shows the average positions of the classification boundaries given by the three methods, trained on $n = 10$ labelled data points and, for the sequentially and hybrid weighted nearest neighbour classifiers, an additional set of $m = 100$ unlabelled points. They are given by $\{x : \mathbb{E}[\hat{p}(x)] = 1/2\}$, $\{x : \mathbb{E}[\hat{s}(x)] = 1/2\}$ and $\{x : \mathbb{E}[\varsigma\hat{p}(x) + (1-\varsigma)\hat{s}(x)] = 1/2\}$, respectively, where the expectations are approximated by averaging over 50 random repetitions. In each repetition, k_1 and k_2 are fixed at the values which minimise test error over the pilot grid for each method.

We see that both the sequentially weighted and hybrid weighted nearest neighbour classifiers yield classification boundaries close to the Bayes boundary in general. They clearly outperform the k_1-nearest neighbour classifier over the low-density area. The above comparison points towards an improvement brought by semi-supervised learning through judicious use of an additional set of unlabelled data.

We next investigate the rates of correct classification achieved by the k_1-nearest neighbour classifier $T_{\boldsymbol{w}_n}$ and our hybrid classifier $\mathscr{T}_{n,m,\varsigma}$, respectively, applied to a test set of size 200. The learning set is made up of a labelled set of size $n = 10$ and an unlabelled set of size $m \in \{5, 10, 100, 500\}$. As before the tuning parameters k_1 and k_2 are set to be the pilot values which minimise test error under each scenario. Note that $T_{\boldsymbol{w}_n}$ depends only on k_1

and does not involve the unlabelled data in its training. The whole experiment is repeated 500 times and the results reported in Table 1.

Table 1: Two-moon data example — correct classification rates (CC) and optimal values of (k_1, k_2) for k_1-nearest neighbour classifier and hybrid weighted nearest neighbour classifier, averaged over 500 replications. Standard deviations are given in parentheses.

	k_1-NN, $T_{\boldsymbol{w}_n}$		hybrid, $\mathscr{T}_{n,m,\varsigma}$		
m	CC	k_1	CC	k_1	k_2
5	0.866 (0.003)	2.232 (1.94)	0.877 (0.002)	2.46 (1.81)	1.92 (1.36)
10	0.867 (0.003)	2.252 (1.90)	0.878 (0.002)	2.54 (1.81)	3.18 (2.69)
100	0.869 (0.002)	2.132 (1.75)	0.879 (0.002)	2.40 (1.72)	32.06 (19.18)
500	0.865 (0.003)	2.092 (1.82)	0.877 (0.002)	2.02 (1.61)	71.66 (43.59)

The hybrid classifier $\mathscr{T}_{n,m,\varsigma}$ shows a stable performance across the four choices of m and is in each case more accurate than the k_1-nearest neighbour classifier, pointing again towards an advantage of our proposed semi-supervised learning approach, even when the number m of unlabelled points is as small as 5.

In the next section we establish asymptotic results to clarify the effects of m and the neighbour weights $\boldsymbol{w}_n$ and $\boldsymbol{w}_m^*$ on the performance of $\mathscr{T}_{n,m,\varsigma}$.

4. Theory

4.1 Assumptions

For a theoretical investigation of $\mathscr{T}_{n,m,\varsigma}$ we adopt the framework laid by Samworth (2012). Suppose that proper densities, f and g say, exist for populations 1 and 2, and that the prior probabilities of these populations are $\pi_f \in (0,1)$ and $\pi_g = 1 - \pi_f$ respectively. Then the Bayes rule assigns $x \in \mathcal{X}$ to the first population if and only if

$$q(x) \equiv \frac{\pi_f f(x)}{\pi_f f(x) + \pi_g g(x)} > \frac{1}{2}. \tag{3}$$

Let $\mathcal{R}$ be a compact d-dimensional manifold in $\mathcal{X}$, which has boundary $\partial\mathcal{R}$ and satisfies

(R1) $\mathcal{S} \equiv \mathcal{R} \cap \{x : q(x) = 1/2\} \neq \emptyset$;

(R2) the restriction of q to $\partial\mathcal{R}$ has nonzero gradient on $\{x \in \partial\mathcal{R} : q(x) = 1/2\}$.

Our focus is on the regret of a classifier $T(\mathcal{L}, x)$, defined to be the misclassification rate of T in excess of the Bayes misclassification rate over $\mathcal{R}$, that is

$$REGRET_{\mathcal{R}}(T) = \mathbb{P}\left(T(\mathcal{L}, \boldsymbol{X}) \neq Y, \boldsymbol{X} \in \mathcal{R}\right) - \mathbb{E}\left[\min\{q(\boldsymbol{X}), 1 - q(\boldsymbol{X})\}; \boldsymbol{X} \in \mathcal{R}\right],$$

where $(\boldsymbol{X}, Y)$ denotes a new observation having the same distribution as $(\boldsymbol{X}_i, Y_i)$, independent of $\mathcal{L}$.

Assume that f, g satisfy the following conditions.

(C1) $\int_{\mathcal{X}} \|x\|^{\delta}\{f(x) + g(x)\}\, dx < \infty$ for some $\delta > 0$.

(C2) For some subset $\mathcal{X}_0$ open in $\mathcal{X}$ and containing $\mathcal{R}$, f, g are four times continuously differentiable on $\mathcal{X}_0$, $\inf_{x\in\mathcal{X}_0} f(x) > 0$, $\inf_{x\in\mathcal{X}_0} g(x) > 0$ and q has nonzero gradient on $\{x \in \mathcal{X}_0 : q(x) = 1/2\}$.

The regularity conditions (R1), (R2), (C1) and (C2) together ensure that the regret over $\mathcal{R}$ is determined essentially by the behaviour of the classifier over $\mathcal{S}$, on which classification becomes most erratic.

Samworth (2012) shows under the above conditions that

$$\min_{\boldsymbol{w}_n} REGRET_{\mathcal{R}}(T_{\boldsymbol{w}_n}) = n^{-4/(d+4)} C_{WNN}\{1 + o(1)\}, \tag{4}$$

where minimisation is over deterministic, non-negative, weight vectors $\boldsymbol{w}_n$ satisfying certain regularity conditions, and $C_{WNN} > 0$ is a constant independent of n. The optimal weights which achieve the minimum (4) have the form

$$w_{n,i} = \begin{cases} \tilde{k}^{-1}\big[1 + d/2 - (d/2)\tilde{k}^{-2/d}\{i^{1+2/d} - (i-1)^{1+2/d}\}\big], & i = 1, \ldots, \tilde{k}, \\ 0, & i = \tilde{k}+1, \ldots, n, \end{cases} \tag{5}$$

for some positive integer $\tilde{k}$ growing at the rate $n^{4/(d+4)}$. Samworth (2012) also shows that the optimal M out of n bagged nearest neighbour classifier and the optimal k-nearest neighbour classifier, both constructed using deterministic weights and trained on $\mathcal{L}_S$, have asymptotic regrets equal to some multiples of the same $n^{-4/(d+4)}$ rate, with the multiplicative factors bigger than C_{WNN} for any finite d. Qiao et al. (2019) propose a big nearest neighbour classifier which achieves the same rate for its regret.

Remark 1 If we remove the assumption of non-negativity on the weights $w_{n,i}$, Samworth (2012) shows that it is possible to improve the optimal order of $REGRET_{\mathcal{R}}(T_{\boldsymbol{w}_n})$ from $O(n^{-4/(d+4)})$ to $O(n^{-8/(d+8)})$ or, more generally, to $O(n^{-4r/(d+4r)})$ under additional smoothness conditions with $r > 1$. However, as commented by Samworth (2012), such accommodation of negative weights requires careful tuning of the dominating bias term and may not be a practically palatable approach to classification. Application of our hybrid procedure might introduce yet more tuning parameters, further complicating the computational algorithm for its practical implementation. For this reason, we do not pursue this approach further in the present work.

We consider in the next sections two special classes of deterministic, non-negative, weights for the construction of the hybrid classifier $\mathscr{T}_{n,m,\varsigma}$, and derive for each class an asymptotic expansion of the regret of $\mathscr{T}_{n,m,\varsigma}$.

For any real sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n \prec b_n$ or $b_n \succ a_n$ if $a_n = o(b_n)$, $a_n \preceq b_n$ or $b_n \succeq a_n$ if $a_n = O(b_n)$, and $a_n \asymp b_n$ if $a_n \preceq b_n$ and $a_n \succeq b_n$.

4.2 Exponentially weighted nearest neighbour

Define, for $\ell > 0$ and any positive integer N,

$$V_{N,\ell,i} = \Big(\frac{1 - e^{-\ell}}{1 - e^{-N\ell}}\Big) e^{-\ell(i-1)}, \quad i = 1, \ldots, N, \tag{6}$$

which constitute a set of deterministic weights decaying exponentially as i increases. For any $\boldsymbol{\ell} = (\ell_1, \ell_2) \in (0, \infty)^2$, denote by $\mathscr{T}^E_{n,m,\varsigma,\boldsymbol{\ell}}$ the hybrid classifier $\mathscr{T}_{n,m,\varsigma}$ with $w_{n,i} = V_{n,\ell_1,i}$ $(i = 1, \ldots, n)$ and $w^*_{m,j} = V_{m,\ell_2,j}$ $(j = 1, \ldots, m)$.

We note that $\mathscr{T}^E_{n,m,\varsigma,\boldsymbol{\ell}}$ shares the same asymptotic properties with any hybrid bagged nearest neighbour classifier constructed using weights given by either (1) or (2), and resample sizes M_S and M_U for $\mathcal{L}_S$ and $\mathcal{L}_U$, respectively, provided that $\boldsymbol{\ell} = (M_S/n, M_U/m) \to (0, 0)$ and $M_S, M_U \to \infty$. The same also holds for hybrid classifiers built by geometric weights $w_{n,i} \asymp (1 - \ell_1)^{i-1}$ and $w^*_{m,j} \asymp (1 - \ell_2)^{j-1}$, provided that $\ell_1, \ell_2 \to 0$ and $n\ell_1, m\ell_2 \to \infty$.

Assume that the unlabelled sample size m increases with the labelled sample size n at a polynomial rate, so that

(M) $n^{a_0} \preceq m \preceq n^{A_0}$ for some constants $A_0 \geq a_0 > 0$.

Let Ω_{d-1} be the $(d-1)$-dimensional volume measure induced on $\mathcal{S}$. Write for brevity $f_\pi = \pi_f f + \pi_g g$. Theorem 1 below establishes an expansion for the regret of $\mathscr{T}^E_{n,m,\varsigma,\boldsymbol{\ell}}$, with the tuning parameter ς set at $1 + \{(m\ell_2)/(n\ell_1)\}^{2/d}$. Its proof is given in the Appendix.

Theorem 1 *Assume that m satisfies (M), f, g satisfy (C1) and (C2), and the compact manifold $\mathcal{R}$ satisfies (R1) and (R2). Let $\varsigma = 1 + \{(m\ell_2)/(n\ell_1)\}^{2/d}$. Then, for any fixed $\epsilon \in \big(0, \{(d+6)A_0\}^{-1}\big)$ and $\boldsymbol{\ell} = (\ell_1, \ell_2)$ satisfying*

$$\ell_1 n^\epsilon + \ell_2 m^\epsilon + (n\ell_1)^{-1} n^{d\epsilon} + (m\ell_2)^{-1} m^{d\epsilon} + m\ell_2 (n\ell_1)^{-1} = O(1), \tag{7}$$

we have

$$REGRET_{\mathcal{R}}\big(\mathscr{T}^E_{n,m,\varsigma,\boldsymbol{\ell}}\big) = \mathscr{R}^V_{\mathcal{R}}\big(\mathscr{T}^E_{n,m,\varsigma,\boldsymbol{\ell}}\big) + \mathscr{R}^B_{\mathcal{R}}\big(\mathscr{T}^E_{n,m,\varsigma,\boldsymbol{\ell}}\big),$$

where

$$\begin{aligned}\mathscr{R}^V_{\mathcal{R}}\big(\mathscr{T}^E_{n,m,\varsigma,\boldsymbol{\ell}}\big) = 8^{-1}\ell_1\{1 + o(1)\} \int_{\mathcal{S}} \|\nabla q(x)\|^{-1} f_\pi(x)\, d\Omega_{d-1}(x) \\ + O\{(n\ell_1)^{-4/d}(m\ell_2)^{-2/d} + \ell_2(n\ell_1)^{-2/d}\}\end{aligned}$$

and

$$\mathscr{R}^B_{\mathcal{R}}\big(\mathscr{T}^E_{n,m,\varsigma,\boldsymbol{\ell}}\big) = O\big(\{\ell_1^2 + \ell_2^2 + (m\ell_2)^{-4/d}\}(n\ell_1)^{-4/d}\big).$$

It can be seen from (A19) and (A20) in the proof that the two components of the regret, $\mathscr{R}^V_{\mathcal{R}}\big(\mathscr{T}^E_{n,m,\varsigma,\boldsymbol{\ell}}\big)$ and $\mathscr{R}^B_{\mathcal{R}}\big(\mathscr{T}^E_{n,m,\varsigma,\boldsymbol{\ell}}\big)$, stem respectively from the variance and squared bias of the hybrid vote proportion $\varsigma\hat{p}(x) + (1 - \varsigma)\hat{s}(x)$ as an estimator of $q(x)$, with $\hat{p}(x) = \sum_{i=1}^n V_{n,\ell_1,i}\mathbf{1}\{Y_{(i)}(x) = 1\}$ and $\hat{s}(x) = \sum_{j=1}^m V_{m,\ell_2,j}\hat{p}(\boldsymbol{Z}_{(j)}(x))$. The expansion (A20) for $\mathscr{R}^B_{\mathcal{R}}\big(\mathscr{T}^E_{n,m,\varsigma,\boldsymbol{\ell}}\big)$ suggests that setting $\varsigma = 1 + \{(m\ell_2)/(n\ell_1)\}^{2/d}$ succeeds in eliminating its leading term. It is noteworthy that the above choice of ς, which has a value bigger than 1, depends only on $\boldsymbol{\ell}$, the dimension d and the sample sizes m, n and can therefore be computed exactly in practice. The above consideration motivates our recommendation of setting $\varsigma = 1 + \{(m\ell_2)/(n\ell_1)\}^{2/d}$. It follows from Theorem 1 that $REGRET_{\mathcal{R}}\big(\mathscr{T}^E_{n,m,\varsigma,\boldsymbol{\ell}}\big) = O(\theta_n)$, where

$$\theta_n = \ell_1 + (n\ell_1)^{-4/d}(m\ell_2)^{-2/d} + \ell_2(n\ell_1)^{-2/d}.$$

We next proceed to minimise θ_n with respect to $\boldsymbol{\ell} = (\ell_1, \ell_2)$ under different asymptotic regimes of m. The following corollary summarises the results. The optimal decaying rates of $\boldsymbol{\ell}$ are deferred to Appendix A.5.

Corollary 2 *Assume the conditions of Theorem 1. Then* $\min\{\theta_n\} \asymp \theta_n^*$*, where the minimum is taken over* (ℓ_1, ℓ_2) *satisfying (7) and the optimal rate* θ_n^* *is specified below across different ranges of* m*.*

(i) $\theta_n^* \asymp n^{-2/(d+2)} m^{-d(1-d\epsilon)/(d+2)}$ *if* $m \preceq n^{2/\{d+4-(d^2+6d+4)\epsilon\}}$.

(ii) $\theta_n^* \asymp \left(n^{4d+4} m^{2d}\right)^{-1/(d^2+6d+4)}$ *if* $n^{2/\{d+4-(d^2+6d+4)\epsilon\}} \preceq m \preceq n^{(d+4)/(d+6)}$.

(iii) $\theta_n^* \asymp n^{-6/(d+6)}$ *if* $m \succeq n^{(d+4)/(d+6)}$.

The change of the optimal regret rate θ_n^* with the unlabelled sample size m is clarified in Figure 3 (solid lines) on the $\log_n$ scale. In general, the convergence rate of θ_n^* increases as the

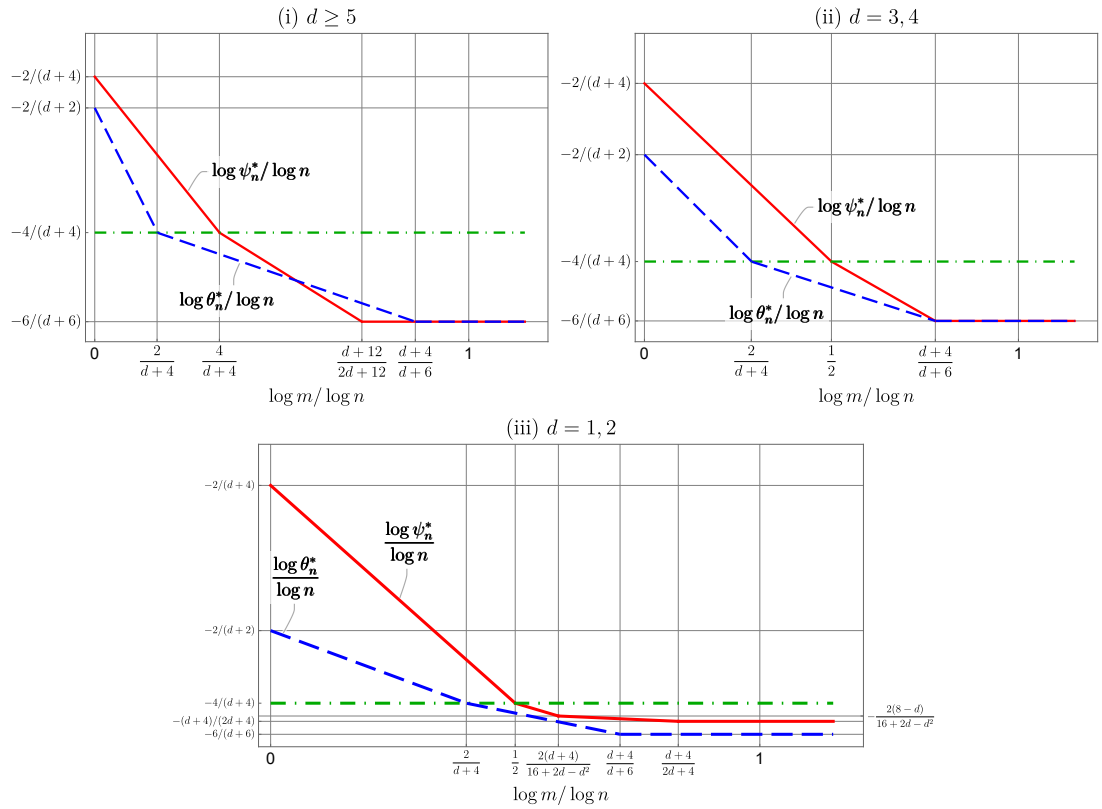

Figure 3: Plots of $\log \theta_n^* / \log n$ (dashed line) and $\log \psi_n^* / \log n$ (solid line) against $\log m / \log n$, with ϵ set at 0.001. Dash-dotted line indicates level $-4/(d+4)$ achieved by optimally weighted nearest neighbour classifier trained on $\mathcal{L}_S$ alone.

divergence rate of m increases, and stabilises at $n^{-6/(d+6)}$ when m grows at a rate faster than $n^{(d+4)/(d+6)}$. As a benchmark, Figure 3 shows also the optimal rate $n^{-4/(d+4)}$ achievable by

training the weighted, bagged or k-nearest neighbour classifier on the labelled set $\mathcal{L}_S$ alone. It is found that the optimal hybrid classifier has a regret converging at a faster rate than $n^{-4/(d+4)}$, that is $\min_{\varsigma,\boldsymbol{\ell}} REGRET_{\mathcal{R}}(\mathscr{T}^E_{n,m,\varsigma,\boldsymbol{\ell}}) \prec \min_{\mathbf{w}_n} REGRET_{\mathcal{R}}(T_{\mathbf{w}_n})$, provided that m has an order exceeding $n^{2/(d+4)}$. As can be seen from the proof of Theorem 1, the hybrid classifier achieves a faster convergence rate essentially by exploiting the prediction made by the sequentially weighted nearest neighbour classifier to correct the bias of the standard classifier $T_{\mathbf{w}_n}$.

Remark 2 (Bootstrap perspective) We may view the hybrid classifier from a novel bootstrap perspective. Let $\{(\boldsymbol{Z}_j, Y^*_j) : j = 1, \ldots, m\}$ be a bootstrap sample generated for the unlabelled set $\mathcal{L}_U$ such that the bootstrap class labels $Y^*_1, \ldots, Y^*_m$ are independent with $\mathbb{P}\big(Y^*_j = 1 \big| \mathcal{L}\big) = \hat{p}(\boldsymbol{Z}_j)$. Setting $\varsigma = 1 + \{(m\ell_2)/(n\ell_1)\}^{2/d}$, the hybrid vote proportion $\varsigma\hat{p}(x) + (1-\varsigma)\hat{s}(x)$ can be rewritten as

$$\hat{p}(x) - \{(m\ell_2)/(n\ell_1)\}^{2/d}\big\{\mathbb{E}[\hat{p}^*(x)|\mathcal{L}] - \hat{p}(x)\big\}, \tag{8}$$

where $\hat{p}^*(x) = \sum_{j=1}^m V_{m,\ell_2,j}\, \mathbf{1}\{Y^*_{(j)}(x) = 1\}$ is the bootstrap analogue of $\hat{p}(x)$ and $Y^*_{(j)}(x)$ denotes the bootstrap class label assigned to $\boldsymbol{Z}_{(j)}(x)$. Viewed as an estimator of $q(x)$, the expression (8) corrects for the bias of the supervised estimator $\hat{p}(x)$ by subtracting a bootstrap bias estimate $\{(m\ell_2)/(n\ell_1)\}^{2/d}\big\{\mathbb{E}[\hat{p}^*(x)|\mathcal{L}] - \hat{p}(x)\big\}$, where the scaling factor $\{(m\ell_2)/(n\ell_1)\}^{2/d}$ adjusts for the difference in size between $\mathcal{L}_S$ and $\mathcal{L}_U$. Since $\mathbb{E}[\hat{p}^*(x)|\mathcal{L}] = \sum_{j=1}^m V_{m,\ell_2,j}\, \hat{p}(\boldsymbol{Z}_{(j)}(x))$ is explicitly available, (8) can be computed exactly without the need for simulating any bootstrap samples in practice.

Remark 3 (Dislabelling strategy) Although our theory requires that the unlabelled sample size m be at least of an order higher than $n^{2/(d+4)}$ if the hybrid classifier is to outperform the optimal $T_{\mathbf{w}_n}$ trained on the labelled sample $\mathcal{L}_S$ alone, in practice the condition can always be met by randomly "dislabelling" a sufficient amount of labelled data points and transferring them to the unlabelled set so as to raise m up to the required order. Indeed, the results of Corollary 2 suggest that if $m = O(n^{2/(d+4)})$, the optimal convergence rate of $REGRET_{\mathcal{R}}(\mathscr{T}^E_{n,m,\varsigma,\boldsymbol{\ell}})$ can be further accelerated to the best possible order $n^{-6/(d+6)}$ by increasing m to at least the order $n^{(d+4)/(d+6)}$. To this end it suffices to transfer, for example, a fixed proportion of data points from $\mathcal{L}_S$ to $\mathcal{L}_U$, thereby increasing the size of the revised $\mathcal{L}_U$ to $m \asymp n$. Since the size of the trimmed $\mathcal{L}_S$ retains the same order as that of the original $\mathcal{L}_S$, the apparent loss of information due to the removal of labels is therefore ratewise inconsequential.

Remark 4 The dislabelling strategy can be applied to our hybrid classifier even under a purely supervised learning setting, that is $m = 0$, to reduce the regret of the optimal $T_{\mathbf{w}_n}$ trained on the original labelled sample. To this end we first create an artificial unlabelled set $\mathcal{L}_U$ by dislabelling a fixed proportion of the original sample. Training the hybrid classifier on the trimmed labelled sample ($\mathcal{L}_S$) and the artificial unlabelled sample ($\mathcal{L}_U$) yields a regret of order $O(n^{-6/(d+6)})$, which converges strictly faster than the standard optimal order $n^{-4/(d+4)}$.

Remark 5 Many of the existing semi-supervised learning approaches make use of unlabelled data to estimate the feature density f_π, knowledge of which may help improve upon supervised learning: see, for example, Lafferty and Wasserman (2007), Sokolovska et al. (2008), Azizyan et al. (2013), Kawakita and Kanamori (2013) and Cannings et al. (2020). To bring out such improvement, a large unlabelled sample size $m \succ n$ is typically required to provide a sufficiently accurate estimate of f_π. By contrast, utilisation of unlabelled data is markedly different in our hybrid semi-supervised procedure. Instead of providing an estimate of f_π, the unlabelled data are used to create a bootstrap-type environment conducive to a correction for the bias of the supervised classifier. Ratewise improvement induced by such bias correction can be achieved with an unlabelled sample size m much smaller than the labelled sample size n.

4.3 Uniformly weighted nearest neighbour

We investigate in this section the effects of our hybrid procedure when applied to standard k-nearest neighbour classifiers, with weights $w_{n,i} = k_1^{-1}\mathbf{1}\{i \le k_1\}$ and $w^*_{m,j} = k_2^{-1}\mathbf{1}\{j \le k_2\}$. Denote by $\mathscr{T}^U_{n,m,\varsigma,\boldsymbol{k}}$ the resulting hybrid weighted nearest neighbour classifier, where $\boldsymbol{k} = (k_1, k_2)$ denotes a pair of positive integers with $k_1 \le n$ and $k_2 \le m$.

The following theorem states an expansion for the regret of $\mathscr{T}^U_{n,m,\varsigma,\boldsymbol{k}}$ analogous to that given in Theorem 1, with the tuning parameter ς set at $1 + \{(k_1/n)/(k_2/m)\}^{2/d}$.

Theorem 3 *Assume the conditions of Theorem 1. Let* $\varsigma = 1 + \{(k_1/n)/(k_2/m)\}^{2/d}$. *Then, for any sufficiently small constant* $\epsilon > 0$ *and* $\boldsymbol{k} = (k_1, k_2)$ *satisfying*

$$(k_1/n)n^\epsilon + (k_2/m)m^\epsilon + n^{4/(d+4)}/k_1 + m^{4/(d+4)}/k_2 + (k_1/n)(k_2/m)^{-1} = O(1), \qquad (9)$$

we have

$$REGRET_{\mathcal{R}}\big(\mathscr{T}^U_{n,m,\varsigma,\boldsymbol{k}}\big) = \mathscr{R}^V_{\mathcal{R}}\big(\mathscr{T}^U_{n,m,\varsigma,\boldsymbol{k}}\big) + \mathscr{R}^B_{\mathcal{R}}\big(\mathscr{T}^U_{n,m,\varsigma,\boldsymbol{k}}\big),$$

where

$$\mathscr{R}^V_{\mathcal{R}}\big(\mathscr{T}^U_{n,m,\varsigma,\boldsymbol{k}}\big) = (4k_1)^{-1}\int_{\mathcal{S}} \|\nabla q(x)\|^{-1} f_\pi(x)\, d\Omega_{d-1}(x)\{1 + o(1)\} + O\big\{(k_1/n)^{6/d} + (k_1/n)^{4/d}(k_2/m)\log n + k_1^{-1/2}(k_1/n)^{4/d}\log n + k_2^{-1/2}(k_1/n)^{2/d}(k_2/m)^{2/d}\log n\big\}$$

and

$$\mathscr{R}^B_{\mathcal{R}}\big(\mathscr{T}^U_{n,m,\varsigma,\boldsymbol{k}}\big) = O\big\{(k_1/n)^{4/d}(k_2/m)^{4/d}\big\}.$$

As with $\mathscr{T}^E_{n,m,\varsigma,\boldsymbol{\ell}}$, setting $\varsigma = 1 + \{(k_1/n)/(k_2/m)\}^{2/d}$ helps eliminate a leading term in the squared bias $\mathscr{R}^B_{\mathcal{R}}\big(\mathscr{T}^U_{n,m,\varsigma,\boldsymbol{k}}\big)$ and yields $REGRET_{\mathcal{R}}\big(\mathscr{T}^U_{n,m,\varsigma,\boldsymbol{k}}\big) = O(\psi_n)$, where

$$\psi_n = k_1^{-1} + (k_1/n)^{6/d} + (k_1/n)^{4/d}(k_2/m)\log n + k_2^{-1/2}(k_1/n)^{2/d}(k_2/m)^{2/d}\log n + (k_1/n)^{4/d}(k_2/m)^{4/d}.$$

Thus, the hybrid classifier $\mathscr{T}^U_{n,m,\varsigma,\boldsymbol{k}}$ also entertains a bootstrap interpretation as discussed in Remark 2, with the scaling factor changed to $\{(k_1/n)/(k_2/m)\}^{2/d}$.

A counterpart of Corollary 2 is given below for the optimal order ψ^*_n of ψ_n, while the corresponding optimal diverging rates of $\boldsymbol{k}$ are presented in Appendix A.6.

Corollary 4 *Assume the conditions of Theorem 3. Then* $\min\{\psi_n\} \asymp \psi_n^*$*, where the minimum is taken over* (k_1, k_2) *satisfying (9) and the optimal rate* ψ_n^* *is specified below across different ranges of* m*.*

(i) For $d \geq 5$, $\psi_n^* \asymp$

(i.1) $n^{-2/(d+4)} m^{-\{1-\epsilon(1-4/d)\}/2} \log n$ *if* $m \preceq \left(n^{2/(d+4)} \log n\right)^{2/\{1-\epsilon(1-4/d)\}}$;

(i.2) $\{n^{-4} m^{-d+\epsilon(d-4)} (\log n)^{2d}\}^{1/(2d+4)}$ *if*

$$(n^{2/(d+4)} \log n)^{2/\{1-\epsilon(1-4/d)\}} \preceq m \preceq (n^{2/(d+4)} \log n)^{2/\{1-\epsilon(d+8)/(d+4)\}};$$

(i.3) $\{nm(\log n)^{-2}\}^{-4/(d+8)}$ *if*

$$(n^{2/(d+4)} \log n)^{2/\{1-\epsilon(d+8)/(d+4)\}} \preceq m \preceq n^{(d+12)/(2d+12)} (\log n)^2;$$

(i.4) $n^{-6/(d+6)}$ *if* $m \succeq n^{(d+12)/(2d+12)} (\log n)^2$.

(ii) For $d = 3, 4$, $\psi_n^* \asymp$

(ii.1) $(nm^2)^{-2/(d+4)} \log n$ *if* $m \preceq n^{1/2} (\log n)^{1+d/4}$;

(ii.2) *, then* $\{n^2 m^{4d/(d+4)} (\log n)^{-d}\}^{-1/(d+2)}$ *if*

$$n^{1/2} (\log n)^{1+d/4} \preceq m \preceq (n \log n)^{(d+4)/(d+6)};$$

(ii.3) $\{n^{8-d} m^d (\log n)^{-2d}\}^{-1/(d+8)}$ *if*

$$(n \log n)^{(d+4)/(d+6)} \preceq m \preceq n^{(d+4)/(d+6)} (\log n)^2;$$

(ii.4) $n^{-6/(d+6)}$ *if* $m \succeq n^{(d+4)/(d+6)} (\log n)^2$.

(iii) For $d = 1, 2$, $\psi_n^* \asymp$

(iii.1) $(nm^2)^{-2/(d+4)} \log n$ *if* $m \preceq n^{1/2} (\log n)^{1+d/4}$;

(iii.2) $\{n^2 m^{4d/(d+4)} (\log n)^{-d}\}^{-1/(d+2)}$ *if*

$$n^{1/2} (\log n)^{1+d/4} \preceq m \preceq (n \log n)^{2(d+4)/(16+2d-d^2)};$$

(iii.3) $\{n^4 m^{d^2/(d+4)} (\log n)^{-d}\}^{-1/(d+4)}$ *if*

$$(n \log n)^{2(d+4)/(16+2d-d^2)} \preceq m \preceq (n \log n)^{(d+4)/(2d+4)};$$

(iii.4) $\{n^{d+4} (\log n)^{-d}\}^{-1/(2d+4)}$ *if* $m \succeq (n \log n)^{(d+4)/(2d+4)}$.

A comparison between Corollaries 2 and 4 suggests that the optimal order of the regret of $\mathscr{T}^U_{n,m,\varsigma,\boldsymbol{k}}$ changes with m in a similar way as that of the regret of $\mathscr{T}^E_{n,m,\varsigma,\boldsymbol{\ell}}$. The optimal regret of $\mathscr{T}^U_{n,m,\varsigma,\boldsymbol{k}}$ converges faster than that of $\mathscr{T}^E_{n,m,\varsigma,\boldsymbol{\ell}}$ if and only if $d \geq 5$ and $\{n^{6d+8}(\log n)^{4(d^2+6d+4)}\}^{1/(d^2+4d+8)} \prec m \prec n^{(d+4)/(d+6)}$. In all cases, the optimal regret of $\mathscr{T}^U_{n,m,\varsigma,\boldsymbol{k}}$ drops to a constant order when m grows at a rate faster than n^c, for some $c < 1$. For $d \geq 3$, this constant order is $n^{-6/(d+6)}$, the same as that of $\mathscr{T}^E_{n,m,\varsigma,\boldsymbol{\ell}}$. For $d \in \{1, 2\}$, the constant order is $\{n^{d+4}(\log n)^{-d}\}^{-1/(2d+4)}$, which is slightly inferior to the corresponding order $n^{-6/(d+6)}$ of $\mathscr{T}^E_{n,m,\varsigma,\boldsymbol{\ell}}$. In case m is not big enough, the dislabelling strategy (Remarks 3, 4) can likewise be applied to $\mathscr{T}^U_{n,m,\varsigma,\boldsymbol{k}}$ to reduce its regret to the constant order, which is strictly smaller than the best order $n^{-4/(d+4)}$ achievable by training a conventional weighted nearest neighbour classifier on $\mathcal{L}_S$ alone. See Figure 3 for a graphical comparison between the optimal regrets of $\mathscr{T}^U_{n,m,\varsigma,\boldsymbol{k}}$ and $\mathscr{T}^E_{n,m,\varsigma,\boldsymbol{\ell}}$.

4.4 Pseudo-labelling by hypothesized class probabilities

Let $q^\dagger$ be a non-negative function on $\mathcal{X}$ and satisfy the regularity conditions:

$$\text{(D)} \quad \int_{\mathcal{X}} \|u\|^\delta q^\dagger(u)\, du < \infty, \quad \inf_{u \in \mathcal{X}_0} q^\dagger(u) > 0, \quad \nabla q^\dagger \neq 0 \text{ on } \{x \in \mathcal{X}_0 : q(x) = 1/2\},$$

where $\delta > 0$ is as specified in (C1). Suppose that the class probabilities $q(\boldsymbol{Z}_j)$ on the unlabelled set $\mathcal{L}_U$ are hypothesized to be $q^\dagger(\boldsymbol{Z}_j)$ $(j = 1, \ldots, m)$, constructed independently of the labelled set $\mathcal{L}_S$. By assigning to $\mathcal{L}_U$ pseudo-labels $q^\dagger(\boldsymbol{Z}_j)$ in place of $\hat{p}(\boldsymbol{Z}_j)$, we may construct a hybrid classifier analogous to $\mathscr{T}^E_{n,m,\varsigma,\boldsymbol{\ell}}$ of the form $\mathcal{T}^E_{n,m,\varsigma,\boldsymbol{\ell}} = 1 + \mathbf{1}\{\hat{\Pi}^E_{HYB,\varsigma}(x) \leq 1/2\}$, where

$$\hat{\Pi}^E_{HYB,\varsigma}(x) = \varsigma \sum_{i=1}^{n} V_{n,\ell_1,i} \mathbf{1}\{Y_{(i)}(x) = 1\} + (1 - \varsigma) \sum_{j=1}^{m} V_{m,\ell_2,j} q^\dagger(\boldsymbol{Z}_{(j)}).$$

Define $\Delta = \int_{\mathcal{X}} |q^\dagger(u) - q(u)| \{f(u) + g(u)\}\, du$, which provides a measure of the mis-specification error of $q^\dagger$.

By setting $(n, \ell_1, \zeta_1) = (\infty, 0, 0)$ in the expansions related to $\hat{s}(x)$ and noting independence between $\sum_{i=1}^n V_{n,\ell_1,i} \mathbf{1}\{Y_{(i)}(x) = 1\}$ and $\sum_{j=1}^m V_{m,\ell_2,j} q^\dagger(\boldsymbol{Z}_{(j)})$, the proof of Theorem 1 can be adapted to show that

$$\begin{aligned}
\mathbb{E}\big[\hat{\Pi}^E_{HYB,\varsigma}(x)\big] = q(x) &+ \{\varsigma(n\ell_1)^{-2/d} + (1-\varsigma)(m\ell_2)^{-2/d}\} \mathcal{V}_d^{-2/d} d^{-1} \Gamma(1 + 2/d) \\
&\times \big\{\nabla q(x)^\top \nabla f_\pi(x) + f_\pi(x) \mathrm{tr}\big(\nabla^2 q(x)\big)/2\big\} f_\pi(x)^{-1-2/d} \\
&+ \varsigma\, O((n\ell_1)^{-2/d}\ell_1 + (n\ell_1)^{-4/d}) + (1-\varsigma) O((m\ell_2)^{-2/d}\ell_2 + (m\ell_2)^{-4/d} + \Delta)
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{Var}\big(\hat{\Pi}^E_{HYB,\varsigma}(x)\big) = \varsigma^2 q(x)\big\{1 - q(x)\big\} \sum_{i=1}^{n} V^2_{n,\ell_1,i} &+ \varsigma^2\, O((n\ell_1)^{-2/d}\ell_1 + (n\ell_1)^{-6/d}) \\
&+ (1-\varsigma)^2 O((m\ell_2)^{-2/d}\ell_2 + (m\ell_2)^{-6/d}).
\end{aligned}$$

The above results suggest setting $\varsigma = \{1-(m\ell_2)^{2/d}(n\ell_1)^{-2/d}\}^{-1}$ in order to eliminate the leading bias term. It then follows that $REGRET_{\mathcal{R}}(\mathcal{T}^E_{n,m,\varsigma,\boldsymbol{\ell}}) = O(\vartheta_n)$, where

$$\vartheta_n = \{1-(m\ell_2)^{2/d}(n\ell_1)^{-2/d}\}^{-2} \times \big\{\ell_1 + (n\ell_1)^{-6/d} + \ell_2(n\ell_1)^{-4/d}(m\ell_2)^{2/d} + (n\ell_1)^{-4/d}(m\ell_2)^{-2/d} + (n\ell_1)^{-4/d}(m\ell_2)^{4/d}\Delta^2\big\}.$$

The following corollary derives the minimum order of $REGRET_{\mathcal{R}}(\mathcal{T}^E_{n,m,\varsigma,\boldsymbol{\ell}})$ by minimising ϑ_n over (ℓ_1,ℓ_2). The optimal choices of (ℓ_1,ℓ_2) are given in Appendix A.7.

Corollary 5 *Assume the conditions (M), (C1), (C2), (R1), (R2) and (D). Let* $\varsigma = \{1-(m\ell_2)^{2/d}(n\ell_1)^{-2/d}\}^{-1}$, *with* $m\ell_2 \neq n\ell_1\{1+o(1)\}$. *Then we have* $\min\{\vartheta_n\} \asymp \vartheta_n^\dagger$, *where the minimum is taken over* (ℓ_1,ℓ_2) *satisfying*

$$\ell_1 n^\epsilon + \ell_2 m^\epsilon + (n\ell_1)^{-1}n^{d\epsilon} + (m\ell_2)^{-1}m^{d\epsilon} = O(1),$$

for any fixed $\epsilon \in \big(0, \{(d+6)A_0\}^{-1}\big)$, *and the optimal rate* $\vartheta_n^\dagger$ *is specified below across different ranges of* m.

(i) For $d \geq 2$, $\vartheta_n^\dagger \asymp$

(i.1) $n^{-4/(d+4)}m^{-2d/(d+4)^2}\{1+m^{6/(d+4)}\Delta^2\}$ *if* $m \preceq n^{(d+4)/(d+6)}$;

(i.2) $n^{-2(d-2)/\{(d+2)(d+6)\}}m^{-4/(d+2)} + \Delta^2$ *if* $m \succeq n^{(d+4)/(d+6)}$.

(ii) For $d = 1$, $\vartheta_n^\dagger \asymp$

(ii.1) $n^{-4/5}m^{-2/25}\{1+m^{6/5}\Delta^2\}$ *if* $m \preceq n^{5/12}$;

(ii.2) $m^{-2} + \Delta^2$ *if* $m \succeq n^{5/12}$.

According to Corollary 5, if the mis-specification error Δ has a sufficiently small order, or more precisely,

$$\Delta \prec \begin{cases} m^{-3/(d+4)} \vee n^{-(d-2)/\{(d+2)(d+6)\}}m^{-2/(d+2)}, & \text{under case (i)},\\ m^{-3/5}, & \text{under case (ii.1)},\\ m^{-1}, & \text{under case (ii.2)}, \end{cases}$$

the optimally tuned $\mathcal{T}^E_{n,m,\varsigma,\boldsymbol{\ell}}$ outperforms the optimally tuned $\mathscr{T}^E_{n,m,\varsigma,\boldsymbol{\ell}}$ in yielding a regret of a strictly smaller order, except when $d \geq 2$ and $m \asymp n^{(d+4)/(d+6)}$, in which case both $\mathcal{T}^E_{n,m,\varsigma,\boldsymbol{\ell}}$ and $\mathscr{T}^E_{n,m,\varsigma,\boldsymbol{\ell}}$ yield optimal regrets of the same order.

Under cases (i.1) or (ii.1) of Corollary 5 when the pseudo-labelled set $\mathcal{L}_U$ is small compared to the labelled set $\mathcal{L}_S$, we have $\varsigma \asymp 1$ and $1-\varsigma \preceq 1$, reflecting a relatively small contribution made by $\mathcal{L}_U$ to $\mathcal{T}^E_{n,m,\varsigma,\boldsymbol{\ell}}$. Under cases (i.2) or (ii.2) when m grows at a sufficiently fast rate, we have $\varsigma \preceq 1$ and $1-\varsigma \asymp 1$, implying that $\mathcal{T}^E_{n,m,\varsigma,\boldsymbol{\ell}}$ becomes dominated by the classifier trained on $\mathcal{L}_U$, whereas the weight given to $\mathcal{L}_S$ shrinks to zero. Thus, the order of the optimal regret of $\mathcal{T}^E_{n,m,\varsigma,\boldsymbol{\ell}}$ continues to decrease without bounds as the diverging rate of m increases. This is in marked contrast to the optimally tuned $\mathscr{T}^E_{n,m,\varsigma,\boldsymbol{\ell}}$, where the weights always satisfy $1 \asymp \varsigma \succeq 1-\varsigma$, so that the contribution made by $\mathcal{L}_S$ never diminishes

asymptotically and hence the optimal regret of $\mathscr{T}^{E}_{n,m,\varsigma,\boldsymbol{\ell}}$ stabilises at a fixed order eventually, no matter how fast m diverges.

Similar results also hold for the case of uniform weights. Consider a hybrid classifier constructed using pseudo-labels $q^{\dagger}(\boldsymbol{Z}_j)$ and uniform weights, namely $\mathcal{T}^{U}_{n,m,\varsigma,\boldsymbol{k}} = 1 + \mathbf{1}\{\hat{\Pi}^{U}_{HYB,\varsigma}(x) \leq 1/2\}$, where

$$\hat{\Pi}^{U}_{HYB,\varsigma}(x) = \varsigma k_1^{-1} \sum_{i=1}^{k_1} \mathbf{1}\{Y_{(i)}(x) = 1\} + (1-\varsigma) k_2^{-1} \sum_{j=1}^{k_2} q^{\dagger}(\boldsymbol{Z}_{(j)}).$$

The proof of Theorem 3 enables us to show that

$$\begin{aligned}\mathbb{E}\big[\hat{\Pi}^{U}_{HYB,\varsigma}(x)\big] = q(x) &+ \{\varsigma(k_1/n)^{2/d} + (1-\varsigma)(k_2/m)^{2/d}\}\mathcal{V}_d^{-2/d} d^{-1}\Gamma(1+2/d) \\ &\times \big\{\nabla q(x)^{\top}\nabla f_{\pi}(x) + f_{\pi}(x)\mathrm{tr}\left(\nabla^2 q(x)\right)/2\big\} f_{\pi}(x)^{-1-2/d} \\ &+ \varsigma\, O((k_1/n)^{4/d}) + (1-\varsigma) O((k_2/m)^{4/d} + \Delta)\end{aligned}$$

and

$$\begin{aligned}\mathrm{Var}\big(\hat{\Pi}^{U}_{HYB,\varsigma}(x)\big) = \varsigma^2 k_1^{-1} q(x)\big\{1-q(x)\big\} &+ \varsigma^2\, O\big\{(k_1/n)^{4/d}(k_1/n + k_1^{-1/2})\log n\big\} \\ &+ (1-\varsigma)^2 O\big\{(k_2/m)^{4/d}(k_2/m + k_2^{-1/2})\log n\big\}.\end{aligned}$$

Setting $\varsigma = \{1 - (k_1/n)^{2/d}(k_2/m)^{-2/d}\}^{-1}$ eliminates the leading bias term of $\hat{\Pi}^{U}_{HYB,\varsigma}(x)$, yielding

$$\begin{aligned}REGRET_{\mathcal{R}}(\mathcal{T}^{U}_{n,m,\varsigma,\boldsymbol{k}}) \asymp \varphi_n = \{1 - (k_1/n)^{2/d}(k_2/m)^{-2/d}\}^{-2} &\\ \times \big\{k_1^{-1} + (k_1/n)^{4/d}(k_1/n + k_2/m + k_1^{-1/2} + k_2^{-1/2})\log n &\\ + (k_1/n)^{8/d} + (k_1/n)^{4/d}((k_2/m)^{4/d} + (k_2/m)^{-4/d}\Delta^2)\big\}.&\end{aligned}$$

Analogous to Corollary 5, the following corollary establishes the minimum order of $REGRET_{\mathcal{R}}(\mathcal{T}^{U}_{n,m,\varsigma,\boldsymbol{k}})$ by minimising φ_n over (k_1, k_2), with the optimal choices of the latter given in Appendix A.8.

Corollary 6 *Assume the conditions (M), (C1), (C2), (R1), (R2) and (D). Let* $\varsigma = \{1 - (k_1/n)^{2/d}(k_2/m)^{-2/d}\}^{-1}$, *with* $mk_1 \neq nk_2\{1+o(1)\}$. *Then we have* $\min\{\varphi_n\} \asymp \varphi_n^{\dagger}$, *where the minimum is taken over* (k_1, k_2) *satisfying*

$$(k_1/n)n^{\epsilon} + (k_2/m)m^{\epsilon} + n^{4/(d+4)}/k_1 + m^{4/(d+4)}/k_2 = O(1),$$

for any sufficiently small constant $\epsilon > 0$, *and the optimal rate* $\varphi_n^{\dagger}$ *is specified below across different ranges of* m.

(i) *For* $d \geq 8$, $\varphi_n^{\dagger} \asymp$

(i.1) $n^{-4/(d+4)}\{m(\log n)^{-2}\}^{-4d/\{(d+4)(d+8)\}}\big[1+\{m(\log n)^{-2}\}^{8/(d+8)}\Delta^2\big]$ *if* $m \preceq n(\log n)^{-d/2}$;

(i.2) $n^{-4/d}\{m^{d-8}(\log n)^{-4d}\}^{-4/(d^2+8d)} + (m/n)^{4/d}\Delta^2$ *if* $n(\log n)^{-d/2} \preceq m \preceq n$;

(i.3) $\{n^{d-8}m^{d+8}(\log n)^{-4d}\}^{-4/(d^2+8d)} + \Delta^2$ *if* $n \preceq m \preceq \{n^4(\log n)^d\}^{(d+4)/(2d+16)}$;

(i.4) $\{n(\log n)^{-2}\}^{-4/(d+8)}m^{-4/(d+4)} + \Delta^2$ *if* $m \succeq \{n^4(\log n)^d\}^{(d+4)/(2d+16)}$.

(ii) *For* $5 \le d \le 7$, $\varphi_n^\dagger \asymp$

(ii.1) $n^{-4/(d+4)}\{m(\log n)^{-2}\}^{-4d/\{(d+4)(d+8)\}}\big[1 + \{m(\log n)^{-2}\}^{4/(d+8)}\Delta^2\big]$ *if*

$$m \preceq n^{(2d+16)/(d+24)}(\log n)^{(d^2+4d+32)/(2d+48)}$$

;

(ii.2) $m^{-6/(d+4)}\log n + \Delta^2$ *if*

$$n^{(2d+16)/(d+24)}(\log n)^{(d^2+4d+32)/(2d+48)} \preceq m \preceq \{n^4(\log n)^d\}^{(d+4)/(2d+16)};$$

(ii.3) $\{n(\log n)^{-2}\}^{-4/(d+8)}m^{-4/(d+4)} + \Delta^2$ *if* $m \succeq \{n^4(\log n)^d\}^{(d+4)/(2d+16)}$.

(iii) *For* $d = 4$, $\varphi_n^\dagger \asymp$

(iii.1) $(n/\log n)^{-1/2}m^{-1/6}\{1 + m^{2/3}(\log n)^{-1}\Delta^2\}$ *if* $m \preceq (n\log n)^{6/7}$;

(iii.2) $m^{-3/4}\log n + \Delta^2$ *if* $(n\log n)^{6/7} \preceq m \preceq n^{4/3}$;

(iii.3) $n^{-1/3}m^{-1/2}\log n + \Delta^2$ *if* $m \succeq n^{4/3}$.

(iv) *For* $2 \le d \le 3$, $\varphi_n^\dagger \asymp$

(iv.1) $\{n^4m^{d/3}(\log n)^{-d}\}^{-1/(d+4)}\{1 + m^{(d+4)/(3d)}(\log n)^{-1}\Delta^2\}$ *if* $m \preceq (n\log n)^{12/(18-d)}$;

(iv.2) $m^{-6/(d+4)}\log n + \Delta^2$ *if* $(n\log n)^{12/(18-d)} \preceq m \preceq (n\log n)^{(d^2+4d)/(4d+8)}$;

(iv.3) $\{n^d(\log n)^{-d-4}\}^{-1/(2d+4)}m^{-4/(d+4)} + \Delta^2$ *if* $m \succeq (n\log n)^{(d^2+4d)/(4d+8)}$.

(v) *For* $d = 1$, $\varphi_n^\dagger \asymp$

(v.1) $n^{-4/5}m^{-1/25}(\log n)^{1/5}\{1 + (m/\log n)\Delta^2\}$ *if* $m \preceq (n\log n)^{5/6}$;

(v.2) $n^{-1/6}m^{-4/5}(\log n)^{5/6} + \Delta^2$ *if* $m \succeq (n\log n)^{5/6}$.

As with $\mathcal{T}^E_{n,m,\varsigma,\boldsymbol{\ell}}$, we see from Corollary 6 that if Δ has a sufficiently small order, the optimally tuned $\mathcal{T}^U_{n,m,\varsigma,\boldsymbol{k}}$ outperforms the optimally tuned $\mathscr{T}^U_{n,m,\varsigma,\boldsymbol{k}}$ in general, except when $1 \le d \le 2$ and $(n\log n)^{(2d+8)/(16+2d-d^2)} \preceq m \preceq (n\log n)^{(d+4)/(2d+4)}$, in which case both $\mathcal{T}^U_{n,m,\varsigma,\boldsymbol{k}}$ and $\mathscr{T}^U_{n,m,\varsigma,\boldsymbol{k}}$ yield optimal regrets of the same order.

A comparison between Corollaries 5 and 6 shows that, provided that the order of Δ is sufficiently small, $\mathcal{T}^E_{n,m,\varsigma,\boldsymbol{\ell}}$ has an optimal regret of an order strictly smaller than that of $\mathcal{T}^U_{n,m,\varsigma,\boldsymbol{k}}$ if and only if $d = 1, 2$ or

$$d \ge 3 \quad \text{and} \quad m \succ \begin{cases} \{n^{(d^2+10d+40)/(4d+24)}(\log n)^{-d-2}\}^{(d+4)/(d+8)}, & d \ge 5, \\ n^{8/5}(\log n)^{-6}, & d = 4, \\ n^{161/144}(\log n)^{-49/16}, & d = 3. \end{cases}$$

5. Numerical examples

5.1 Simulation study I

We conduct a simulation study to compare the performance of our hybrid semi-supervised classifier $\mathscr{T}^{U}_{n,m,\varsigma,\boldsymbol{k}}$ (Section 4.3) with two existing supervised classifiers, namely the standard k-nearest neighbour classifier and the optimally weighted nearest neighbour classifier (Samworth, 2012) constructed using weights (5). Throughout the study the mixing coefficient ς is set to be $1 + \big\{(k_1/n)/(k_2/m)\big\}^{2/d}$.

Our study covers a variety of scenarios, under dimensions $d \in \{5, 10, 15, 20\}$, labelled sample sizes $n \in \{50, 100, 200, 500\}$ and unlabelled sample sizes $m \in \{0, 2, 10, 40, 100, 200\}$. Under each scenario, both the labelled and unlabelled samples are generated from two distinct populations such that one half of each sample belongs to class 1 and the other half to class 2. We consider four different model settings for the two densities, f and g, of the two populations. They are detailed below, with $W(a, b)$ denoting the Weibull distribution with shape parameter a and scale parameter b, $L(a, b)$ denoting the Laplace distribution with mean a and scale parameter b, and $N(\boldsymbol{\mu}, \Sigma)$ denoting the, possibly multivariate, normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ.

1. $f(x_1, \ldots, x_d) = \prod_{i=1}^{d} f_0(x_i)$ and $g(x_1, \ldots, x_d) = \prod_{i=1}^{d} g_0(x_i)$, where f_0 and g_0 are the $L(0, 1)$ and $N(1, 1)$ densities, respectively.

2. f and g are densities of the d-variate normal mixtures $\frac{1}{2}N(0\boldsymbol{e}_d, \Sigma) + \frac{1}{2}N(3\boldsymbol{e}_d, 2\Sigma)$ and $\frac{1}{2}N(\frac{3}{2}\boldsymbol{e}_d, \Sigma) + \frac{1}{2}N(\frac{9}{2}\boldsymbol{e}_d, 2\Sigma)$, respectively, where $\boldsymbol{e}_d = [1, \ldots, 1]^\top$ and Σ is set to be the $d \times d$ Toeplitz matrix with its $(1, j)$th entry given by $(0.6)^{j-1}$.

3. $f(x_1, \ldots, x_d) = \prod_{i=1}^{d} f_0(x_i)$ and $g(x_1, \ldots, x_d) = \prod_{i=1}^{d} g_0(x_i)$, where f_0 is the density of the Weibull mixture $\frac{1}{2}W(1, 1) + \frac{1}{2}W(3, \frac{1}{2})$ and g_0 is the density of the Laplace-Weibull mixture $\frac{1}{2}L(\frac{1}{2}, \frac{1}{2}) + \frac{1}{2}W(5, 3)$.

4. $f(x_1, \ldots, x_d) = \prod_{i=1}^{d} f_0(x_i)$ and $g(x_1, \ldots, x_d) = \prod_{i=1}^{\lfloor d/2 \rfloor} f_0(x_i) \times \prod_{i=\lfloor d/2 \rfloor + 1}^{d} g_0(x_i)$, where f_0 is the Cauchy density with location parameter $3/2$ and scale parameter 1, and g_0 is the $L(0, 2)$ density.

In addition to the formal hybrid classifier $\mathscr{T}^{U}_{n,m,\varsigma,\boldsymbol{k}}$, we include in our study a revised version, denoted $\mathscr{T}^{U*}_{n,m,\varsigma,\boldsymbol{k}}$, motivated by the dislabelling strategy described in Remark 3, for the cases with $m < n$. Specifically, we construct $\mathscr{T}^{U*}_{n,m,\varsigma,\boldsymbol{k}}$ by the following procedure:

Step 1. Select randomly a subsample of size $(n - m)/2$ without replacement from the labelled set, dislabel them and merge them with the unlabelled set, resulting in a pair of revised labelled and unlabelled sets of the same size $(n + m)/2$.

Step 2. Train $\mathscr{T}^{U}_{n,m,\varsigma,\boldsymbol{k}}$ on the revised labelled and unlabelled sets obtained at *Step 1*, and predict the test points.

Step 3. Repeat *Steps 1* and *2* J times to obtain J predictions for each test point.

Step 4. Determine the final prediction for each test point by a majority vote among its J predictions.

Our empirical results show that fixing J to a number between 20 and 100 is sufficient for a stable prediction made by $\mathscr{T}^{U*}_{n,m,\varsigma,\boldsymbol{k}}$. We fix $J = 20$ in the study.

In each scenario and for each replication of simulation, we find the rate of misclassifying a set of 1000 test data points under each pilot combination of tuning parameters, and extract the lowest rate to exemplify the best possible performance achievable by each method under optimal tuning. The whole process is replicated 1000 times. The regret of each method under each scenario is then approximated by the average of the lowest misclassification rates over the 1000 replications less the Bayes risk, which is approximated by the rate of misclassifying a random sample of 10^7 data points using the Bayes rule. To assess the improvement made by the revised hybrid classifier relative to a particular method, we calculate the ratio regret(other method)/regret($\mathscr{T}^{U*}_{n,m,\varsigma,\boldsymbol{k}}$).

Figures 4–7 show the relative improvements of $\mathscr{T}^{U*}_{n,m,\varsigma,\boldsymbol{k}}$ over the other three methods, namely the k-nearest neighbour classifier, the optimally weighted nearest neighbour classifier and the original hybrid classifier $\mathscr{T}^{U}_{n,m,\varsigma,\boldsymbol{k}}$, under labelled sample sizes $n = 50$, 100, 200 and 500, respectively. Note that the original hybrid classifier is undefined in the absence of unlabelled data and is therefore omitted from the case $m = 0$.

We see that the revised hybrid classifier $\mathscr{T}^{U*}_{n,m,\varsigma,\boldsymbol{k}}$ succeeds in reducing regret of nearest neighbour classification under all settings, even in the case $m = 0$ where no additional unlabelled data are available. The reduction is biggest under setting 1 and smallest under setting 4. The improvement made by the optimally weighted nearest neighbour classifier over the standard k-nearest neighbour classifier is less significant than that made by both of our hybrid classifiers $\mathscr{T}^{U*}_{n,m,\varsigma,\boldsymbol{k}}$ and $\mathscr{T}^{U}_{n,m,\varsigma,\boldsymbol{k}}$. Indeed, under setting 3, the optimally weighted nearest neighbour classifier has the biggest regret among all the methods. This corroborates our theoretical findings that the hybrid classifier enjoys a faster convergence rate for its optimal regret than that of the k-nearest neighbour classifier, while optimal weighting fails to improve upon the latter ratewise. We also observe, by comparing $\mathscr{T}^{U*}_{n,m,\varsigma,\boldsymbol{k}}$ against $\mathscr{T}^{U}_{n,m,\varsigma,\boldsymbol{k}}$, that the dislabelling strategy is effective in boosting the performance of the original hybrid method in general.

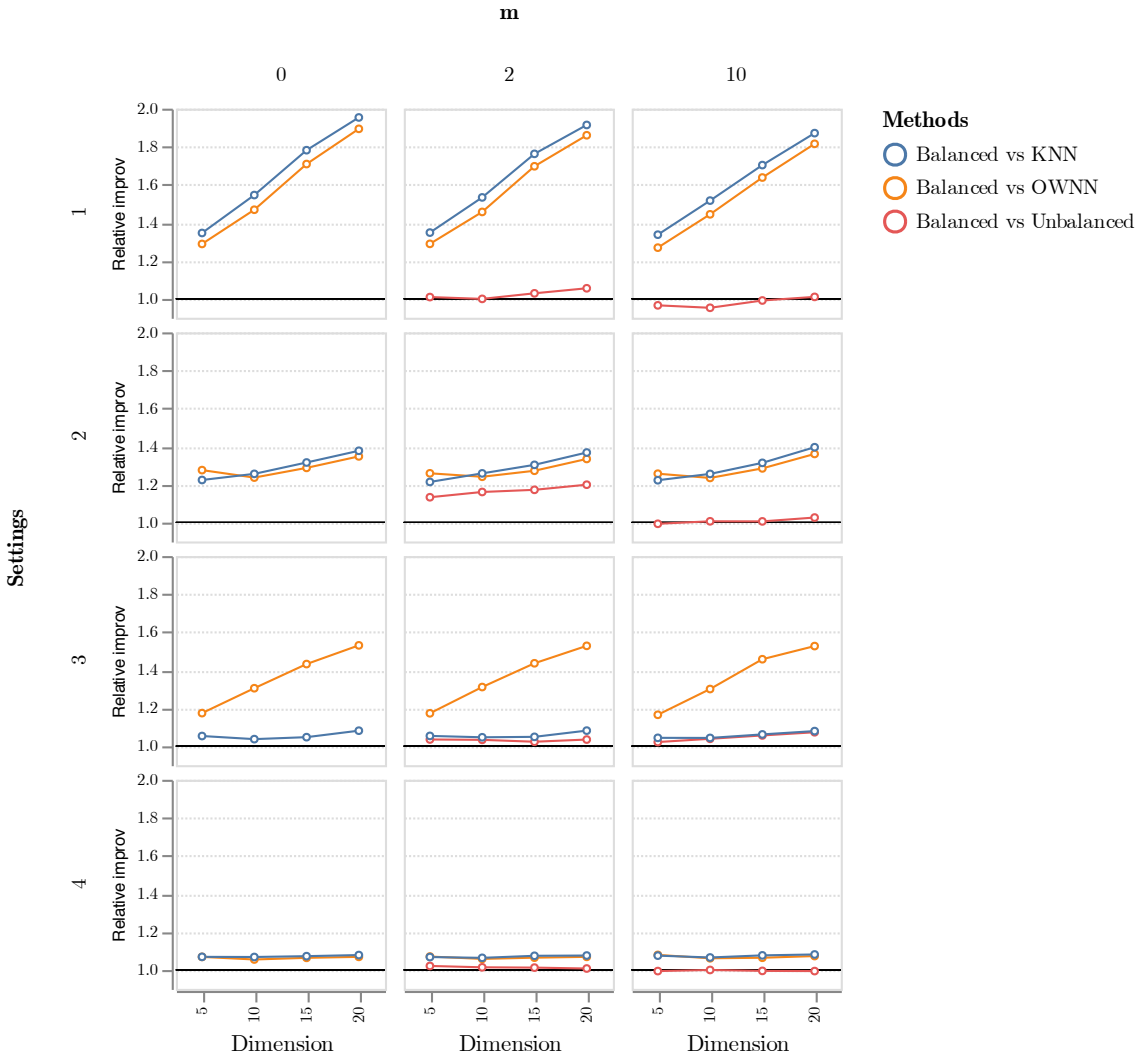

Figure 4: Relative improvement in regret of revised hybrid classifier (Balanced) vs original hybrid classifier (Unbalanced), k-nearest neighbour classifier (KNN) and optimally weighted nearest neighbour classifier (OWNN), for $n = 50$.

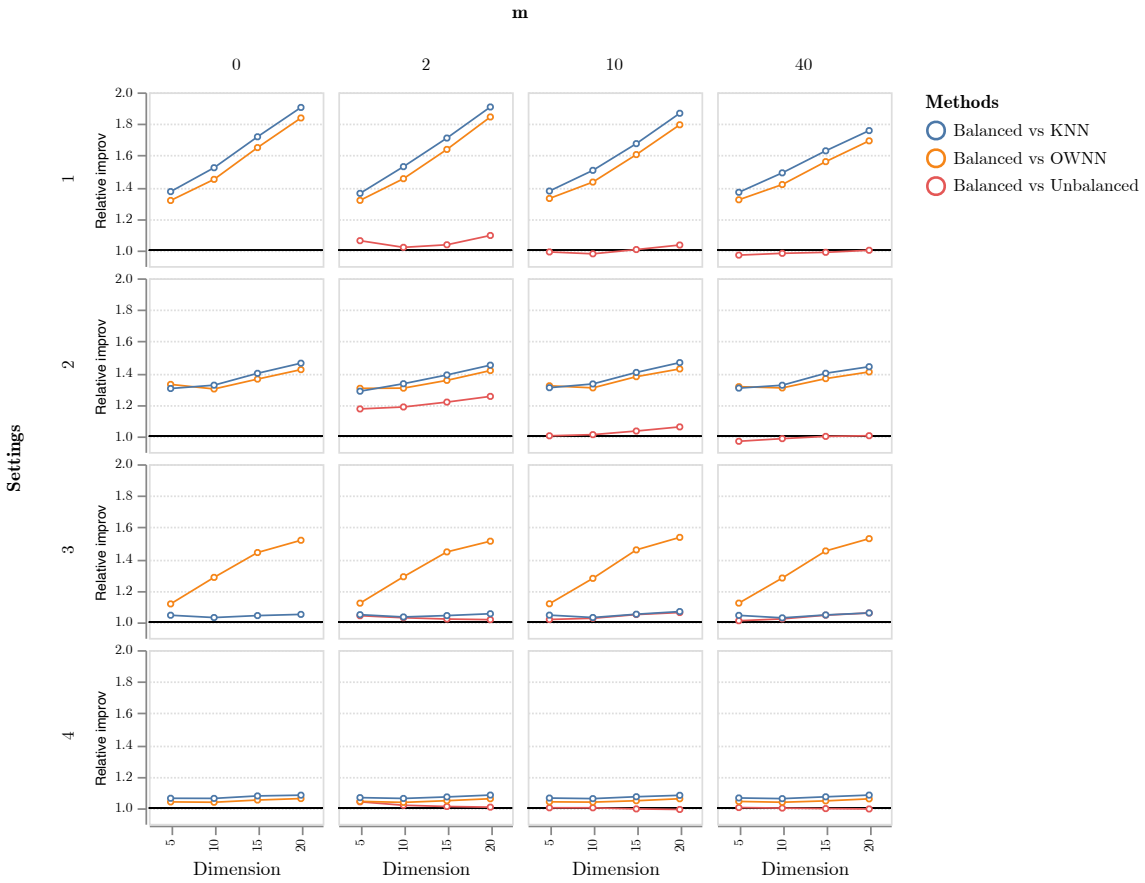

Figure 5: Relative improvement in regret of revised hybrid classifier (Balanced) vs original hybrid classifier (Unbalanced), k-nearest neighbour classifier (KNN) and optimally weighted nearest neighbour classifier (OWNN), for $n = 100$.

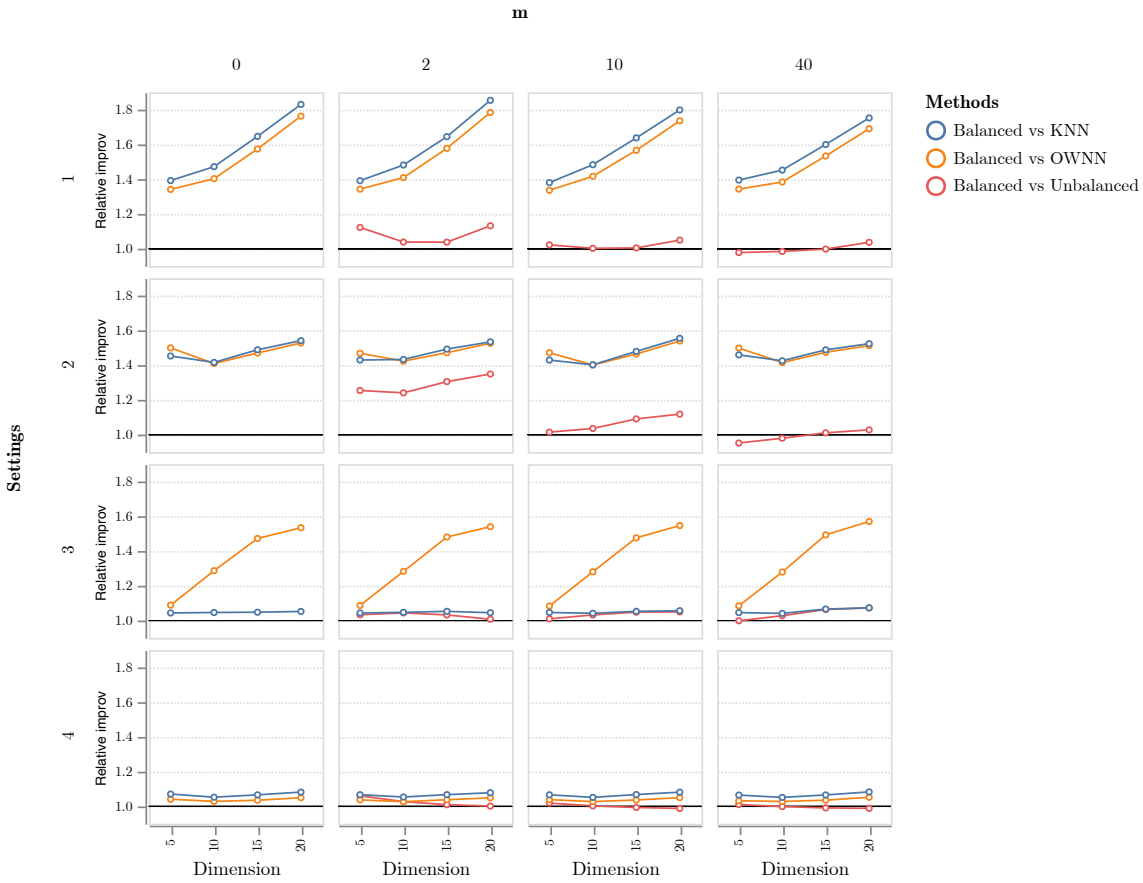

Figure 6: Relative improvement in regret of revised hybrid classifier (Balanced) vs original hybrid classifier (Unbalanced), k-nearest neighbour classifier (KNN) and optimally weighted nearest neighbour classifier (OWNN), for $n = 200$.

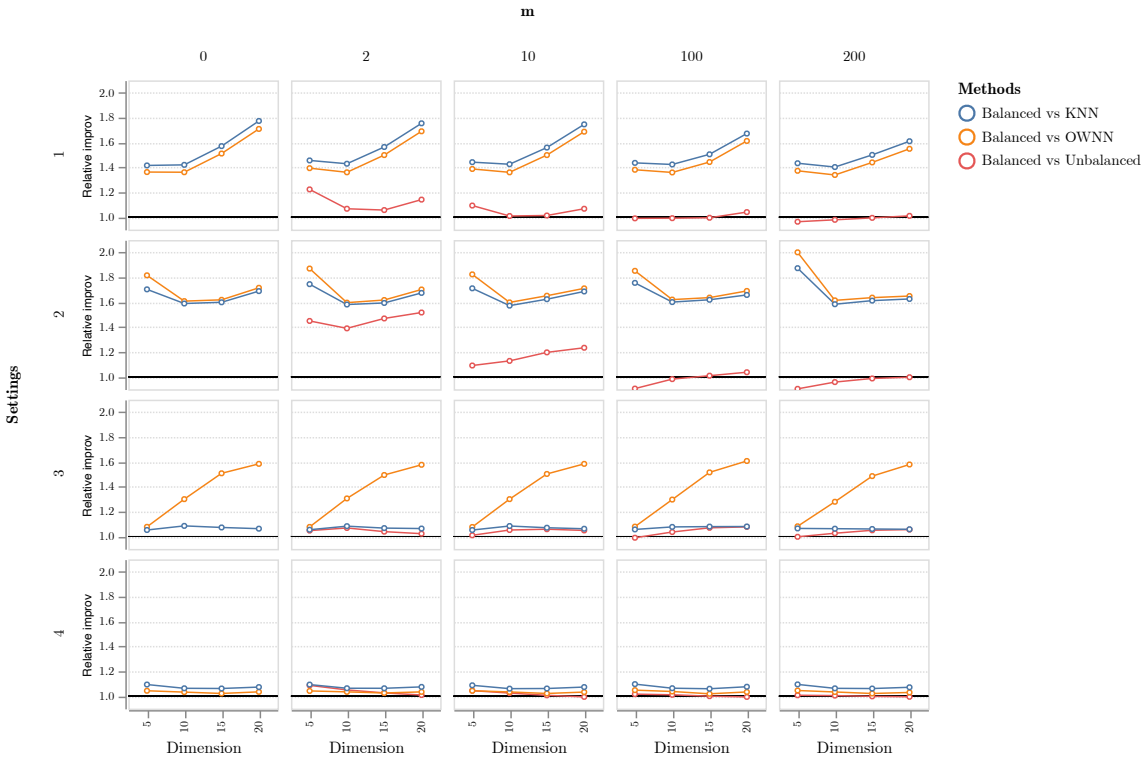

Figure 7: Relative improvement in regret of revised hybrid classifier (Balanced) vs original hybrid classifier (Unbalanced), k-nearest neighbour classifier (KNN) and optimally weighted nearest neighbour classifier (OWNN), for $n = 500$.

5.2 Simulation study II

In practice, the numbers of nearest neighbours need to be determined empirically for the various nearest neighbour classification methods. We have conducted a second simulation study to compare the revised hybrid classifier $\mathscr{T}^{U*}_{n,m,\varsigma,\boldsymbol{k}}$ against the k-nearest neighbour and optimally weighted nearest neighbour classifiers, with their respective tuning parameters $\boldsymbol{k}$ and k fixed by 5-fold cross-validation. In particular, the optimally weighted nearest neighbour classifier is tuned by the modified 5-fold cross-validation algorithm described in Samworth (2012). We set in this study $n \in \{50, 100\}$, $d \in \{5, 10, 15, 20\}$ and $m = 0$. The hybrid classifier $\mathscr{T}^{U*}_{n,m,\varsigma,\boldsymbol{k}}$ is constructed using $J = 100$ iterations of the dislabelling step.

Figure 8 shows the improvements in regret of $\mathscr{T}^{U*}_{n,m,\varsigma,\boldsymbol{k}}$ relative to the other two methods. Generally speaking, $\mathscr{T}^{U*}_{n,m,\varsigma,\boldsymbol{k}}$ has a smaller regret than that of the optimally weighted nearest neighbour classifier, by a margin that grows with the dimension d. The same holds for its improvement over the k-nearest neighbour classifier, except under setting 3 where the latter has the smallest regret among the three methods. The above findings suggest that cross-validation works satisfactorily to bring forth the theoretical advantages of our proposed hybrid method in practical applications.

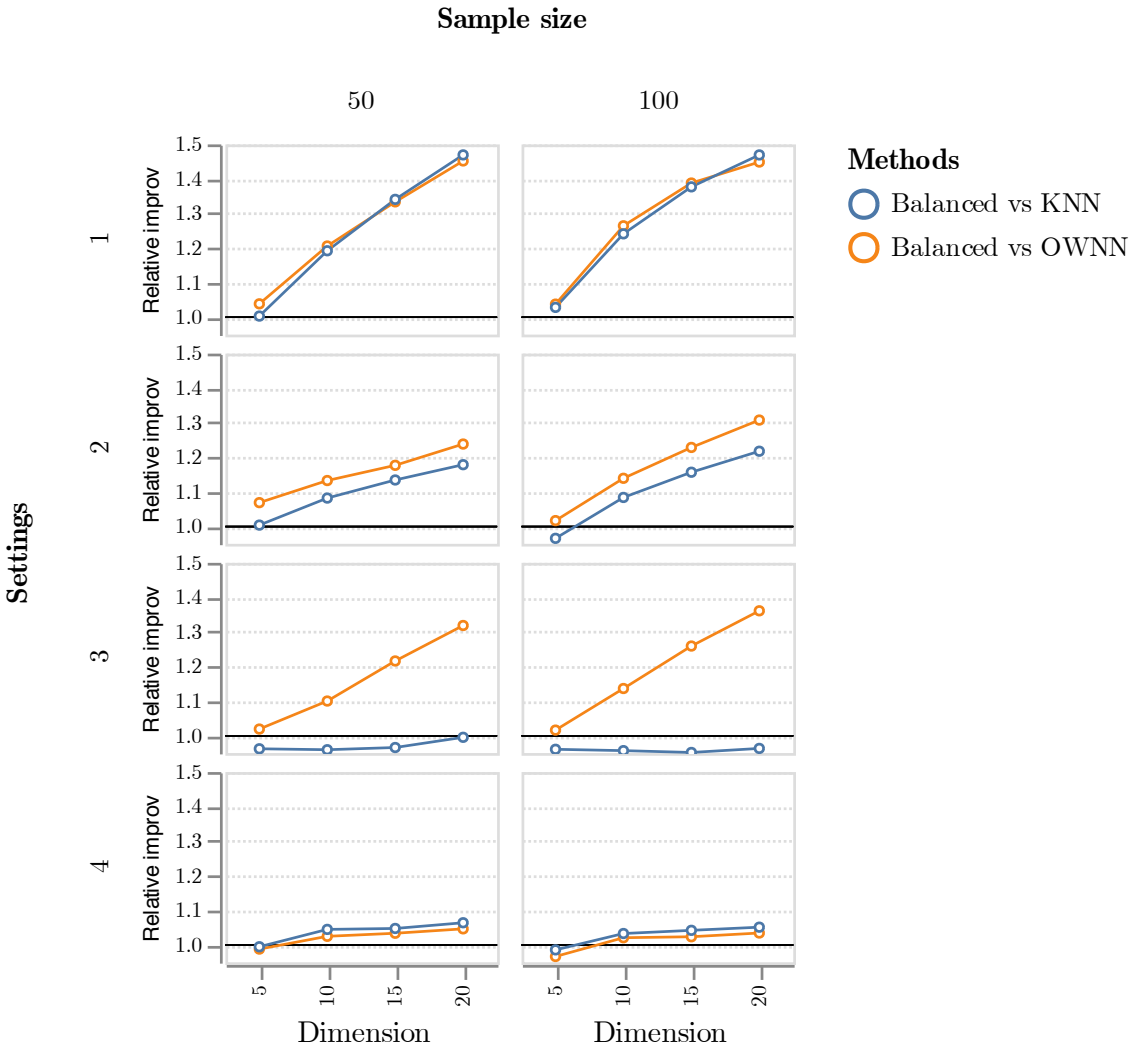

Figure 8: Relative improvement in regret of revised hybrid classifier (Balanced) vs k-nearest neighbour classifier (KNN) and optimally weighted nearest neighbour classifier (OWNN), with tuning parameters fixed by 5-fold cross-validation.

5.3 Real data applications

Six real data sets are selected from the UCI repository (Lichman, 2013) to benchmark the revised hybrid weighted nearest neighbour classifier $\mathscr{T}^{U*}_{n,m,\varsigma,\boldsymbol{k}}$ against the k-nearest neighbour and optimally weighted nearest neighbour classifiers. They are, respectively,

1. Post-operative patient data set, used for determining where patients in a postoperative recovery area should be sent to next;
2. Ecoli data set, used for predicting protein localisation sites in Gram-Negative bacteria;
3. Wisconsin breast cancer database, used to diagnose breast cancer;
4. Yeast data, used for predicting cellular localisation sites of proteins;
5. Musk data, used for predicting whether new molecules will be musks or non-musks;
6. MAGIC gamma telescope data, use to simulate registration of high energy gamma particles in an atmospheric Cherenkov telescope.

To find the best tuning parameters for the proposed method, we fixed the k_1 and k_2 values to the largest possible values for each fold in the 5-fold cross-validation and used a scan search to find the k-nearest neighbours for each point in the labelled set and the unlabelled set. This approach allows us to obtain predictions for the out-of-sample points of the cross validation process for any values of k_1 and k_2 that are less than the largest possible values without rerunning the k-nearest neighbour search. We select the best-performing k_1 and k_2 from this cross-validation to predict the test set.

We modify the Post-operative data set by removing observations with missing values, and the Ecoli and Yeast data sets by combining rare classes having small numbers of instances. For the musk data we removed `molecule_name` and `conformation_name`. The data points in each set are all labelled, scaled and then randomly assigned to a learning set and a test set with probabilities 0.3 and 0.7, respectively. Tuning parameters $\boldsymbol{k}$ and k are selected by cross-validation. Except for the MAGIC data which we set the repetition to 100, for each method and each data set, the average of misclassification rates over 1000 repetitions of train-test assignments is reported in Table 2.

Our results show that the hybrid classifier $\mathscr{T}^{U*}_{n,m,\varsigma,\boldsymbol{k}}$ gives a lower average misclassification rate than the other two methods, although the margin of improvement appears not very significant.

6. Concluding remarks

We have introduced a hybrid weighted nearest neighbour classifier $\mathscr{T}_{n,m,\varsigma}$ for semi-supervised learning. It linearly combines a standard weighted nearest neighbour classifier, trained on a labelled sample of size n, and a sequentially weighted nearest neighbour classifier, trained on a pseudo-labelled sample of size m. When optimally tuned, $\mathscr{T}_{n,m,\varsigma}$ has a regret converging at a faster rate than that of the optimally weighted nearest neighbour classifier, and hence that of any k-nearest neighbour classifier, provided that m exceeds an order of the form n^c, for some $c \in (0, 1)$ depending on the dimension d and the type of weights chosen for

Table 2: Misclassification rates, averaged over 1000 train-test assignments, of revised hybrid classifier (Balanced), optimally weighted nearest neighbour classifier (OWNN) and k-nearest neighbour classifier (KNN), trained on three selected UCI repository data sets. Columns n, d, and K show the labelled sample size, the number of predictors and the number of response classes, respectively. Standard deviations are given in parentheses.

Data set	n	d	K	Balanced	OWNN	KNN
Post Operative	85	8	2	0.310 (0.047)	0.377 (0.064)	0.314 (0.053)
Ecoli	336	7	6	0.171 (0.023)	0.181 (0.027)	0.174 (0.022)
Cancer	569	30	2	0.056 (0.011)	0.057 (0.011)	0.056 (0.011)
Yeast	1484	9	7	0.444 (0.014)	0.449 (0.022)	0.453 (0.014)
Musk	476	168	2	0.190 (0.026)	0.197 (0.027)	0.192 (0.023)
MAGIC	19020	10	2	0.199 (0.004)	0.206 (0.003)	0.206 (0.006)

the classifier. We have also proposed a dislabelling strategy to revise $\mathscr{T}_{n,m,\varsigma}$. The regret of the resulting revised classifier achieves the optimal rate, which is faster than that of the optimally weighted nearest neighbour classifier, under any unlabelled sample size m including the purely supervised case $m = 0$. The above theoretical findings are supported by empirical results obtained from simulations and real data applications. We have also shown that standard cross-validation provides a practically reliable approach to tuning of neighbour weights for the hybrid classifier.

References

Azizyan, M., Singh, A. and Wasserman, L. (2013). Density-sensitive semisupervised inference. *Ann. Statist.*, **41**, 751–771.

Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A. and Raffel, C. (2019). MixMatch: a holistic approach to semi-supervised learning. arXiv:1905.02249 [cs.LG].

Belongie, S., Malik, J. and Puzicha, J. (2002). Shape matching and object recognition using shape contexts. *IEEE Trans. Pattern Anal. Mach. Intell.*, **24**, 509–522.

Berrett, T.B. and Samworth, R.J. (2019). Efficient two-sample functional estimation and the super-oracle phenomenon. arXiv:1904.09347 [math.ST].

Biau, G., Cérou, F. and Guyader, A. (2010). On the rate of convergence of the bagged nearest neighbour estimate. *J. Mach. Learn. Res.*, **11**, 687–712.

Biau, G. and Devroye, L. (2015). *Lectures on the Nearest Neighbor Method*. Springer International Publishing Switzerland.

Blum, A. and Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. In *Proceedings of the 11th Annual Conference on Computational Learning Theory*, pp.92–100.

Cannings, T.I., Berrett, T.B. and Samworth, R.J. (2020). Local nearest neighbour classification with applications to semi-supervised learning. *Ann. Statist.*, **48**, 1789–1814.

Dudani, S.A. (1976). The distance-weighted k-nearest-neighbour rule. *IEEE Transactions on Systems, Man and Cybernetics*, **SMC-6**, 325–327.

Fix, E. and Hodges, J.L., Jr. (1951). Discriminatory analysis, nonparametric discrimination: consistency properties. Report no. 4, Project no. 21-49-004, USAF School of Aviation Medicine, Randolph Field, Texas.

Gadat, S., Klein, T. and Marteau, C. (2016). Classification in general finite dimensional spaces with the k-nearest neighbor rule. *Ann. Statist.*, **44**, 982–1009.

Grandvalet, Y. and Bengio, Y. (2004). Semi-supervised learning by entropy minimization. In *Proceedings of the 17th International Conference on Neural Information Processing Systems*, pp.529–536.

Hall, P. and Samworth, R.J. (2005). Properties of bagged nearest neighbour classifiers. *J. Roy. Statist. Soc. Ser. B*, **67**, 363–379.

Joachims, T. (1999). Transductive inference for text classification using support vector machines. In *Proceedings of the 16th International Conference on Machine Learning*, pp.200–209.

Kawakita, M. and Kanamori, T. (2013). Semi-supervised learning with density-ratio estimation. *Mach. Learn.*, **91**, 189–209.

Lafferty, J. and Wasserman, L. (2007). Statistical analysis of semi-supervised regression. Advances in Neural Information Processing Systems **20** (NIPS 2007), (eds., J. Platt, D. Koller, Y. Singer and S. Roweis).

Li, L., Darden, T.A., Weinberg, C.R., Levine, A.J. and Pedersen, L.G. (2001). Gene assessment and sample classification for gene expression data using a genetic algorithm/k-nearest neighbor method. *Comb. Chem. High Throughput Screen.*, **4**, 727–739. doi: 10.2174/1386207013330733.

Lichman, M. (2013). UCI Machine Learning Repository Irvine, CA: University of California, School of Information and Computer Science. Available at http://archive.ics.uci.edu/ml.

Liu, Z., Zhao, X., Zou, J. and Xu, H. (2013). A semi-supervised approach based on k-nearest neighbor. *Journal of Software*, **8**, 768–775.

Nigam, K., McCallum, A.K., Thrun, S. and Mitchell, T. (2000). Text classification from labeled and unlabeled documents using EM. *Machine Learning*, **39**, 103–134.

Qiao, X., Duan, J. and Cheng, G. (2019). Rates of convergence for large-scale nearest neighbor classification. In *Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*.

Rebuffi, S.-A., Ehrhardt, S., Han, K., Vedaldi, A. and Zisserman, A. (2020). Semi-supervised learning with scarce annotations. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp.3294–3302, doi: 10.1109/CVPRW50498.2020.00389.

Samworth, R.J. (2012). Optimal weighted nearest neighbour classifiers. *Ann. Statist.*, **40**, 2733–2763.

Sokolovska, N., Cappé, O. and Yvon, F. (2008). The asymptotics of semi-supervised learning in discriminative probabilistic models. In *Proceedings of the 25th international conference on machine learning*, pp.984–991.

Soleymani, M. and Lee, S.M.S. (2014). Sequential combination of weighted and nonparametric bagging for classification. *Biometrika*, **101**, 491–498.

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

Tu, E., Zhang, Y., Zhu, L., Yang, J. and Kasabov, N. (2016). A graph-based semi-supervised k nearest-neighbor method for nonlinear manifold distributed data classification. *Information Sciences*, **367-368**, 673–688.

van Engelen, J.E. and Hoos, H.H. (2019). A survey on semi- supervised learning. *Machine Learning*.**109**, 373–440.

Vapnik, V. (1998). *Statistical Learning Theory*. Wiley: New York.

Wajeed, M.A. and Adilakshmi, T. (2011). Semi-supervised text classification using enhanced KNN algorithm. In *Proceedings of 2011 World Congress on Information and Communication Technologies*, pp.138–142.

Wang, Y., Xu, X., Zhao, H. and Hua, Z. (2010). Semi-supervised learning based on nearest neighbor rule and cut edges. *Knowledge-Based Systems*, **23**, 547–554.

Zhu, X. and Goldberg, A.B. (2009). *Introduction to Semi-supervised Learning*. Morgan & Claypool Publishers: San Rafael.

Appendix

A.1 Proof of Theorem 1

Recall that $\hat{p}(x) = \sum_{i=1}^{n} V_{n,\ell_1,i}\mathbf{1}\{Y_{(i)}(x) = 1\}$ and $\hat{s}(x) = \sum_{j=1}^{m} V_{m,\ell_2,j}\,\hat{p}(\boldsymbol{Z}_{(j)}(x))$. Define $\hat{\Pi}_{HYB,\varsigma}(x) = \varsigma\hat{p}(x) + (1-\varsigma)\hat{s}(x)$. Then, for all $x \in \mathcal{X}$, $\mathscr{T}^{E}_{n,m,\varsigma,\boldsymbol{\ell}}(\mathcal{L}, x) = 1$ if and only if $\hat{\Pi}_{HYB,\varsigma}(x) > 1/2$.

Define, for $z, z_1, z_2 \in \mathbb{R}^d$ and $a, b > 0$, $p_a(z) = \mathbb{P}\big(\|\boldsymbol{X} - z\| < a\big)$,

$$\begin{aligned}
\varrho_{a,b}(z_1, z_2) &= \mathbb{P}\big(\|\boldsymbol{X} - z_1\| < a,\, \|\boldsymbol{X} - z_2\| < b\big),\\
\varrho_2(a, b, z_1, z_2) &= \mathbb{P}\big(\|\boldsymbol{X} - z_1\| < a,\, \|\boldsymbol{X} - z_2\| \geq b\big) = p_a(z_1) - \varrho_{a,b}(z_1, z_2),\\
\varrho_3(a, b, z_1, z_2) &= \mathbb{P}\big(\|\boldsymbol{X} - z_1\| \geq a,\, \|\boldsymbol{X} - z_2\| < b\big) = p_b(z_2) - \varrho_{a,b}(z_1, z_2),\\
\varrho_4(a, b, z_1, z_2) &= 1 - p_a(z_1) - p_b(z_2) + \varrho_{a,b}(z_1, z_2).
\end{aligned}$$

Define, for $\eta \in \big(0, \epsilon\min\{4/d, 1\}\big)$,

$$\mathcal{X}_{n,\eta} = \big\{x + t\|\nabla q(x)\|^{-1}\nabla q(x) : |t| < n^{-\eta/(4d)},\ \inf_{y\in\mathcal{S}}\|x - y\| < n^{-\eta/(4d)},\ q(x) = 1/2\big\}.$$

Define, for any $z, z_1, z_2 \in \mathcal{X}$, any $\alpha, \ell > 0$, any positive integer N, and any smooth functions h on $\mathcal{X}_{n,\eta}$ and H on $\mathcal{X}_{n,\eta} \times \mathcal{X}_{n,\eta}$ satisfying $H(z_1, z_2) = H(z_2, z_1)$,

$$M_{\alpha,N,\ell}(z; h) = \sum_{i=1}^{N} V_{N,\ell,i}^{\alpha}\mathbb{E}\big[h(\boldsymbol{Z}'_{(i)}(z))\big] \tag{A1}$$

and

$$\begin{aligned}
\tilde{M}_{N,\ell}(z_1, z_2; H) &= \sum_{i,i'=1}^{N} V_{N,\ell,i}V_{N,\ell,i'}\Big\{\mathbb{E}\big[H(\boldsymbol{Z}'_{(i)}(z_1), \boldsymbol{Z}'_{(i)}(z_1)); \boldsymbol{Z}'_{(i)}(z_1) = \boldsymbol{Z}'_{(i')}(z_2)\big]\\
&\quad + \mathbb{E}\big[H(\boldsymbol{Z}'_{(i)}(z_1), \boldsymbol{Z}'_{(i')}(z_2)); \boldsymbol{Z}'_{(i)}(z_1) \neq \boldsymbol{Z}'_{(i')}(z_2)\big]\Big\}\\
&= I_{N,\ell}(z_1, z_2; H) + II_{N,\ell}(z_1, z_2; H),
\end{aligned} \tag{A2}$$

where $(\boldsymbol{Z}'_1, \ldots, \boldsymbol{Z}'_N)$ denotes a random sample drawn from the density f_π.

We first establish expansions for (A1) and (A2) under the assumption $\ell N^{\epsilon} + (N\ell)^{-1}N^{d\epsilon} = O(1)$.

Let $\mathcal{V}_d = \pi^{d/2}\Gamma(1 + d/2)^{-1}$ denote the volume of the unit ball in $\mathbb{R}^d$. Writing $\zeta = (N\ell\mathcal{V}_d)^{-1/d}$, we have

$$\begin{aligned}
M_{\alpha,N,\ell}(z; h) &= \Big(\frac{1 - e^{-\ell}}{1 - e^{-N\ell}}\Big)^{\alpha} N\int\{h(z + t) - h(z)\}f_\pi(z + t)\\
&\quad \times \big\{1 - (1 - e^{-\alpha\ell})p_{\|t\|}(z)\big\}^{N-1}\,dt + h(z)\sum_{i=1}^{N} V_{N,\ell,i}^{\alpha}\\
&= \Big(\frac{1 - e^{-\ell}}{1 - e^{-N\ell}}\Big)^{\alpha} N\zeta^d\int\{h(z + \zeta u) - h(z)\}f_\pi(z + \zeta u)\\
&\quad \times e^{-N\ell\alpha p_{\zeta\|u\|}(z)}\{1 + O(\ell)\}\,du + h(z)\sum_{i=1}^{N} V_{N,\ell,i}^{\alpha}.
\end{aligned} \tag{A3}$$

Similarly, writing $H_0(z) = H(z,z)$, we may express the expectations in (A2) in integral forms such that

$$I_{N,\ell}(z_1,z_2;H) = \sum_{i,i'=1}^{N} V_{N,\ell,i}V_{N,\ell,i'} \int H_0(z_1+t) g^I_{N,i,i',z_1,z_2}(t)\,dt \tag{A4}$$

and

$$II_{N,\ell}(z_1,z_2;H) = \sum_{i,i'=1}^{N} V_{N,\ell,i}V_{N,\ell,i'} \iint H(z_1+t,z_2+s) g^{II}_{N,i,i',z_1,z_2}(t,s)\,dt\,ds, \tag{A5}$$

where

$$\begin{aligned}
g^I_{N,i,i',z_1,z_2}(t) &\triangleq \frac{\partial}{\partial t}\,\mathbb{P}\big(\boldsymbol{Z}'_{(i)}(z_1) \le z_1+t; \boldsymbol{Z}'_{(i)}(z_1) = \boldsymbol{Z}'_{(i')}(z_2)\big) \\
&= N f_\pi(z_1+t) \\
&\quad \times \sum_{k=\max\{0,i+i'-N-1\}}^{\min\{i,i'\}-1} \mathcal{M}_{N-1,\|t\|,\|z_1-z_2+t\|,z_1,z_2}(k, i-1-k, i'-1-k, N-i-i'+k+1),
\end{aligned}$$

$$\begin{aligned}
g^{II}_{N,i,i',z_1,z_2}(t,s) &\triangleq \frac{\partial^2}{\partial s\,\partial t}\,\mathbb{P}\big(\boldsymbol{Z}'_{(i)}(z_1) \le z_1+t, \boldsymbol{Z}'_{(i')}(z_2) \le z_2+s; \boldsymbol{Z}'_{(i)}(z_1) \ne \boldsymbol{Z}'_{(i')}(z_2)\big) \\
&= N(N-1) f_\pi(z_1+t) f_\pi(z_2+s) \Big[\mathbf{1}\{\|z_1-z_2+t\| > \|s\|,\, \|z_2-z_1+s\| > \|t\|\} \\
&\quad \times \sum_{k=\max\{0,i+i'-N\}}^{\min\{N-1,i,i'\}-1} \mathcal{M}_{N-2,\|t\|,\|s\|,z_1,z_2}(k, i-1-k, i'-1-k, N-i-i'+k) \\
&\quad + \mathbf{1}\{\|z_1-z_2+t\| < \|s\|,\, \|z_2-z_1+s\| < \|t\|\} \\
&\quad \times \sum_{k=\max\{0,i+i'-N-2\}}^{\min\{i+i'-2,i,i'\}-2} \mathcal{M}_{N-2,\|t\|,\|s\|,z_1,z_2}(k, i-2-k, i'-2-k, N-i-i'+k+2) \\
&\quad + \mathbf{1}\{\|z_1-z_2+t\| > \|s\|,\, \|z_2-z_1+s\| < \|t\|\} \\
&\quad \times \sum_{k=\max\{0,i+i'-N-1,i'-N+1\}}^{\min\{i-2,i'-1\}} \mathcal{M}_{N-2,\|t\|,\|s\|,z_1,z_2}(k, i-2-k, i'-1-k, N-i-i'+k+1) \\
&\quad + \mathbf{1}\{\|z_1-z_2+t\| < \|s\|,\, \|z_2-z_1+s\| > \|t\|\} \\
&\quad \times \sum_{k=\max\{0,i+i'-N-1,i-N+1\}}^{\min\{i-1,i'-2\}} \mathcal{M}_{N-2,\|t\|,\|s\|,z_1,z_2}(k, i-1-k, i'-2-k, N-i-i'+k+1) \Big]
\end{aligned}$$

and $\mathcal{M}_{N,\alpha,\beta,z_1,z_2}(\cdot)$ denotes the multinomial $\big(N; \varrho_{\alpha,\beta}(z_1,z_2), p_\alpha(z_1) - \varrho_{\alpha,\beta}(z_1,z_2), p_\beta(z_2) - \varrho_{\alpha,\beta}(z_1,z_2), 1 - p_\alpha(z_1) - p_\beta(z_2) + \varrho_{\alpha,\beta}(z_1,z_2)\big)$ mass function, for any $\alpha,\beta > 0$, positive integer N and $z_1, z_2 \in \mathbb{R}^d$. We note, in particular, that

$$\iint g^{II}_{N,i,i',z_1,z_2}(t,s)\,dt\,ds = \mathbb{P}\big(\boldsymbol{Z}'_{(i)}(z_1) \ne \boldsymbol{Z}'_{(i')}(z_2)\big) = 1 - \int g^I_{N,i,i',z_1,z_2}(t)\,dt.$$

Interchanging summations and integrations, and setting $t = \zeta u$ and $s = \zeta v$ in the integrals, (A4) and (A5) reduce to

$$\begin{aligned} I_{N,\ell}(z_1, z_2; H) &= H_0(z_1) \sum_{i,i'=1}^{N} V_{N,\ell,i} V_{N,\ell,i'} \int g^I_{N,i,i',z_1,z_2}(t)\, dt \\ &\quad + \Big(\frac{1-e^{-\ell}}{1-e^{-N\ell}}\Big)^2 N\zeta^d \int \{H_0(z_1+\zeta u) - H_0(z_1)\} \\ &\quad \times f_\pi(z_1+\zeta u) e^{-N\ell\{p_{\zeta\|u\|}(z_1) + p_{\|z_1-z_2+\zeta u\|}(z_2)\}} \big\{1 + O(\ell)\big\}\, du \end{aligned} \tag{A6}$$

and

$$\begin{aligned} &II_{N,\ell}(z_1, z_2; H) \\ &= H(z_1, z_2) \sum_{i,i'=1}^{N} V_{N,\ell,i} V_{N,\ell,i'} \Big\{1 - \int g^I_{N,i,i',z_1,z_2}(t)\, dt\Big\} \\ &\quad + N(N-1)\Big(\frac{1-e^{-\ell}}{1-e^{-N\ell}}\Big)^2 \zeta^{2d} \iint \{H(z_1+\zeta u, z_2+\zeta v) - H(z_1, z_2)\} \\ &\quad \times f_\pi(z_1+\zeta u) f_\pi(z_2+\zeta v) e^{-N\ell\{p_{\zeta\|u\|}(z_1) + p_{\zeta\|v\|}(z_2)\}} \big\{1 + O(\ell)\big\}\, du\, dv, \end{aligned} \tag{A7}$$

respectively.

Let $\zeta_1 = (n\ell_1 \mathcal{V}_d)^{-1/d}$ and $\zeta_2 = (m\ell_2 \mathcal{V}_d)^{-1/d}$. Define, for any $\alpha \geq 0$, any $x \in \mathcal{X}_{n,\eta}$ and any smooth function h on $\mathcal{X}_{n,\eta}$,

$$\begin{aligned} \iota_{1,\alpha}(x) &= \int \|u\|^\alpha e^{-n\ell_1 p_{\zeta_1\|u\|}(x)}\, du, \\ \mathcal{J}_1(x; h) &= \zeta_1^{-2} \int \{h(x+\zeta_1 u) - h(x)\} f_\pi(x+\zeta_1 u) e^{-n\ell_1 p_{\zeta_1\|u\|}(x)}\, du. \end{aligned}$$

Define $\iota_{2,\alpha}(x)$ and $\mathcal{J}_2(x; h)$ similarly, with (n, ℓ_1, ζ_1) replaced by (m, ℓ_2, ζ_2). Taylor expanding h and f_π about x, and using the fact that $p_a(x) = f_\pi(x)\mathcal{V}_d a^d\{1 + O(a^2)\}$ as $a \downarrow 0$, uniformly over $x \in \mathcal{X}_{n,\eta}$, it can be shown that $\iota_{j,\alpha}(\cdot) \asymp 1$ and $\mathcal{J}_j(\cdot) \asymp 1$, $j = 1, 2$. By repeated use of (A3) with $\alpha = 1$, we have

$$\begin{aligned} \mathbb{E}\big[\hat{s}(x)\big] &= M_{1,m,\ell_2}\big(x; M_{1,n,\ell_1}(\cdot; q)\big) \\ &= q(x) + \Big(\frac{1-e^{-\ell_1}}{1-e^{-n\ell_1}}\Big) n\zeta_1^{d+2} \mathcal{J}_1(x; q) + \Big(\frac{1-e^{-\ell_2}}{1-e^{-m\ell_2}}\Big) m\zeta_2^{d+2} \mathcal{J}_2(x; q) \\ &\quad + mn\Big(\frac{1-e^{-\ell_2}}{1-e^{-m\ell_2}}\Big)\Big(\frac{1-e^{-\ell_1}}{1-e^{-n\ell_1}}\Big) \zeta_1^{d+2} \zeta_2^{d+2} \mathcal{J}_2\big(x; \mathcal{J}_1(\cdot; q)\big) + O(\zeta_1^2 \ell_1 + \zeta_2^2 \ell_2) \end{aligned} \tag{A8}$$

uniformly over $x \in \mathcal{X}_{n,\eta}$.

For any real-valued function q on $\mathcal{X}$, denote by $q^{\otimes}$ the function mapping $(x,y) \in \mathcal{X}\times\mathcal{X}$ to $q(x)\mathbf{1}\{x=y\} + q(x)q(y)\mathbf{1}\{x\neq y\}$. Using (A2), (A6) and (A7), we have

$$\begin{aligned}
\mathbb{E}\big[\hat{s}(x)^2\big] &= \tilde{M}_{m,\ell_2}\big(x,x;\tilde{M}_{n,\ell_1}(\cdot,\cdot;q^{\otimes})\big) \\
&= \tilde{M}_{n,\ell_1}(x,x;q^{\otimes}) + m(m-1)\Big(\frac{1-e^{-\ell_2}}{1-e^{-m\ell_2}}\Big)^2\zeta_2^{2d}\iint f_\pi(x+\zeta_2 u)f_\pi(x+\zeta_2 v) \\
&\quad\times\{\tilde{M}_{n,\ell_1}(x+\zeta_2 u, x+\zeta_2 v;q^{\otimes}) - \tilde{M}_{n,\ell_1}(x,x;q^{\otimes})\} \\
&\quad\times e^{-m\ell_2\{p_{\zeta_2\|u\|}(x)+p_{\zeta_2\|v\|}(x)\}}\,du\,dv + O(\ell_2\zeta_2^2). \qquad \text{(A9)}
\end{aligned}$$

Again using (A6) and (A7), and Taylor expanding $q^{\otimes}$, we have, for $u\neq v$,

$$\begin{aligned}
&\tilde{M}_{n,\ell_1}(x+\zeta_2 u, x+\zeta_2 v;q^{\otimes}) \\
&= \Big(\frac{1-e^{-\ell_1}}{1-e^{-n\ell_1}}\Big)^2 n\zeta_1^d\int\{q(x+\zeta_2 u+\zeta_1 w) - q(x+\zeta_2 u)\}f_\pi(x+\zeta_2 u+\zeta_1 w) \\
&\quad\times e^{-n\ell_1\{p_{\zeta_1\|w\|}(x+\zeta_2 u)+p_{\|\zeta_2(u-v)+\zeta_1 w\|}(x+\zeta_2 v)\}}\big\{1+O(\ell_1)\big\}\,dw \\
&\quad+ n(n-1)\Big(\frac{1-e^{-\ell_1}}{1-e^{-n\ell_1}}\Big)^2\zeta_1^{2d}\Big\{\zeta_1^4\mathcal{J}_1(x+\zeta_2 u;q)\mathcal{J}_1(x+\zeta_2 v;q) \\
&\quad+\zeta_1^2 q(x+\zeta_2 u)\mathcal{J}_1(x+\zeta_2 v;q)\int f_\pi(x+\zeta_2 u+\zeta_1 w)e^{-n\ell_1 p_{\zeta_1\|w\|}(x+\zeta_2 u)}\,dw \\
&\quad+\zeta_1^2 q(x+\zeta_2 v)\mathcal{J}_1(x+\zeta_2 u;q)\int f_\pi(x+\zeta_2 v+\zeta_1 t)e^{-n\ell_1 p_{\zeta_1\|t\|}(x+\zeta_2 v)}\,dt\Big\} \\
&\quad+ q(x+\zeta_2 u)q(x+\zeta_2 v) + q(x+\zeta_2 u)\big\{1-q(x+\zeta_2 v)\big\}I_{n,\ell_1}(x+\zeta_2 u, x+\zeta_2 v;1) \\
&\quad+ O(\ell_1\zeta_1^2) \qquad \text{(A10)}
\end{aligned}$$

and

$$\begin{aligned}
&\tilde{M}_{n,\ell_1}(x,x;q^{\otimes}) \\
&= q(x)^2 + q(x)\big\{1-q(x)\big\}\sum_{i=1}^n V_{n,\ell_1,i}^2 + n(n-1)\Big(\frac{1-e^{-\ell_1}}{1-e^{-n\ell_1}}\Big)^2\zeta_1^{2d} \\
&\quad\times\Big\{\zeta_1^4\mathcal{J}_1(x;q)^2 + 2\zeta_1^2 q(x)\mathcal{J}_1(x;q)\int f_\pi(x+\zeta_1 w)e^{-n\ell_1 p_{\zeta_1\|w\|}(x)}\,dw\Big\} + O(\ell_1\zeta_1^2). \qquad \text{(A11)}
\end{aligned}$$

Substituting (A10) and (A11) into (A9), and using the expansion

$$\begin{aligned}
I_{n,\ell_1}(x+\zeta_2 u, x+\zeta_2 v;1) &= \Big(\frac{1-e^{-\ell_1}}{1-e^{-n\ell_1}}\Big)^2 n\zeta_1^d\int f_\pi(x+\zeta_2 u+\zeta_1 w) \\
&\quad\times e^{-n\ell_1\{p_{\zeta_1\|w\|}(x+\zeta_2 u)+p_{\|\zeta_2(u-v)+\zeta_1 w\|}(x+\zeta_2 v)\}}\big\{1+O(\ell_1)\big\}\,dw,
\end{aligned}$$

we have

$$\begin{aligned}
&\mathbb{E}\big[\hat{s}(x)^2\big] \\
&= q(x)^2 + q(x)\big\{1-q(x)\big\}\sum_{i=1}^{n} V_{n,\ell_1,i}^2 + n(n-1)\Big(\frac{1-e^{-\ell_1}}{1-e^{-n\ell_1}}\Big)^2\zeta_1^{2d} \\
&\quad\times\Big\{\zeta_1^4\mathcal{J}_1(x;q)^2 + 2\zeta_1^2 q(x)\mathcal{J}_1(x;q)\int f_\pi(x+\zeta_1 w)e^{-n\ell_1 p_{\zeta_1\|w\|}(x)}\,dw\Big\} \\
&\quad + m(m-1)\Big(\frac{1-e^{-\ell_2}}{1-e^{-m\ell_2}}\Big)^2\zeta_2^{2d} \\
&\quad\times\Big\{\zeta_2^4\mathcal{J}_2(x;q)^2 + 2\zeta_2^2 q(x)\mathcal{J}_2(x;q)\int f_\pi(x+\zeta_2 u)e^{-m\ell_2 p_{\zeta_2\|u\|}(x)}\,du\Big\} \\
&\quad + 2nm(n-1)(m-1)\Big(\frac{1-e^{-\ell_1}}{1-e^{-n\ell_1}}\Big)^2\Big(\frac{1-e^{-\ell_2}}{1-e^{-m\ell_2}}\Big)^2\zeta_1^{2d+2}\zeta_2^{2d+2} \\
&\quad\times\Big\{q(x)f_\pi(x)\iota_{1,0}(x)\mathcal{J}_2\big(x;\mathcal{J}_1(\cdot;q)\big) + \mathcal{J}_1(x;q)\mathcal{J}_2\big(x;qf_\pi\iota_{1,0}\big)\Big\}f_\pi(x)\iota_{2,0}(x) \\
&\quad + O(\ell_1^2+\ell_1\zeta_2^2+\ell_2\zeta_2^2+\zeta_1^2\zeta_2^4). \qquad \text{(A12)}
\end{aligned}$$

Noting the identity $\int \xi(\|u\|)\,du = d\mathcal{V}_d\int_0^\infty \xi(r)r^{d-1}dr$, for any integrable function $\xi(\|u\|)$ over $u\in\mathbb{R}^d$, and the expansion

$$p_\alpha(x) = \mathcal{V}_d\alpha^d f_\pi(x) + \alpha^{d+2}(2d+4)^{-1}\mathcal{V}_d\mathrm{tr}\left(\nabla^2 f_\pi(x)\right) + O(\alpha^{d+4}), \quad \alpha>0,$$

it can be shown that, for $j=1,2$ and $\alpha\geq 0$,

$$\begin{aligned}
\iota_{j,\alpha}(x) &= \mathcal{V}_d\Gamma(1+\alpha/d)f_\pi(x)^{-1-\alpha/d} + O(\zeta_j^2), \\
\mathcal{V}_d^{-1}f_\pi(x)\iota_{j,0}(x) &= 1-\zeta_j^2(2d+4)^{-1}\Gamma(2+2/d)f_\pi(x)^{-1-2/d}\mathrm{tr}\left(\nabla^2 f_\pi(x)\right) + O(\zeta_j^4), \\
\mathcal{J}_j(x;h) &= d^{-1}\big\{\nabla h(x)^\top\nabla f_\pi(x) + f_\pi(x)\mathrm{tr}\left(\nabla^2 h(x)\right)/2\big\}\iota_{j,2}(x) + O(\zeta_j^2).
\end{aligned}$$

It then follows from (A12) and (A8) that, uniformly over $x\in\mathcal{X}_{n,\eta}$,

$$\begin{aligned}
\mathrm{Var}\big(\hat{s}(x)\big) &= q(x)\big\{1-q(x)\big\}\sum_{i=1}^{n} V_{n,\ell_1,i}^2 + 2\mathcal{V}_d^{-1}q(x)\sum_{j=1}^{2}\zeta_j^4\mathcal{J}_j(x;q) \\
&\quad\times\Big[\zeta_j^{-2}\big\{\mathcal{V}_d^{-1}f_\pi(x)\iota_{j,0}(x)-1\big\} + d^{-1}\mathcal{V}_d^{-1}\mathrm{tr}\left(\nabla^2 f_\pi(x)\right)\iota_{j,2}(x)/2\Big] \\
&\quad + O(\ell_1^2+\ell_1\zeta_2^2+\ell_2\zeta_2^2+\zeta_2^6+n^{-1}\zeta_1^4) \\
&= q(x)\big\{1-q(x)\big\}\sum_{i=1}^{n} V_{n,\ell_1,i}^2 + O(\ell_1^2+\ell_1\zeta_2^2+\ell_2\zeta_2^2+\zeta_2^6+n^{-1}\zeta_1^4). \qquad \text{(A13)}
\end{aligned}$$

Analogous results can be obtained for the exponentially weighted vote proportion $\hat{p}(x) = \sum_{i=1}^n V_{n,\ell,i}\mathbf{1}\{Y_{(i)}(x)=1\}$ by applying the same arguments to $\mathbb{E}\big[\hat{p}(x)\big] = M_{1,n,\ell_1}(x;q)$ and $\mathbb{E}\big[\hat{p}(x)^2\big] = \tilde{M}_{n,\ell_1}(x,x;q^{\otimes})$, yielding

$$\mathbb{E}\big[\hat{p}(x)\big] = q(x) + \Big(\frac{1-e^{-\ell_1}}{1-e^{-n\ell_1}}\Big)n\zeta_1^{d+2}\mathcal{J}_1(x;q) + O(\ell_1\zeta_1^2), \qquad \text{(A14)}$$

$$\mathrm{Var}\big(\hat{p}(x)\big) = q(x)\big\{1-q(x)\big\}\sum_{i=1}^{n} V_{n,\ell_1,i}^2 + O(\ell_1\zeta_1^2+\zeta_1^6). \qquad \text{(A15)}$$

Noting the identity $\mathbb{E}\big[\hat{p}(x)\hat{s}(x)\big] = M_{1,m,\ell_2}\big(x; \tilde{M}_{n,\ell_1}(x,\cdot;q^{\otimes})\big)$ and using (A3), (A11) and (A10) (with v set to 0), the above proof can be adapted to show that

$$\begin{aligned}\mathbb{E}\big[\hat{p}(x)\hat{s}(x)\big] &= q(x)^2 + q(x)\big\{1-q(x)\big\}\sum_{i=1}^{n} V_{n,\ell_1,i}^2 + n(n-1)\Big(\frac{1-e^{-\ell_1}}{1-e^{-n\ell_1}}\Big)^2\zeta_1^{2d+2} \\ &\quad\times\Big\{\zeta_1^2\mathcal{J}_1(x;q)^2 + 2q(x)\mathcal{J}_1(x;q)\int f_\pi(x+\zeta_1 w)e^{-n\ell_1 p_{\zeta_1\|w\|}(x)}\,dw\Big\} \\ &\quad + m\Big(\frac{1-e^{-\ell_2}}{1-e^{-m\ell_2}}\Big)\zeta_2^d\Big[\zeta_2^2 q(x)\mathcal{J}_2(x;q) + n(n-1)\Big(\frac{1-e^{-\ell_1}}{1-e^{-n\ell_1}}\Big)^2\zeta_1^{2d+2}\zeta_2^2 \\ &\quad\times\Big\{q(x)f_\pi(x)\iota_{1,0}(x)\mathcal{J}_2\big(x;\mathcal{J}_1(\cdot;q)\big) + \mathcal{J}_1(x;q)\mathcal{J}_2\big(x;qf_\pi\iota_{1,0}\big)\Big\}\Big] \\ &\quad + O(\ell_1^2 + \ell_1\zeta_1^2 + \ell_1\zeta_2^2 + \ell_2\zeta_2^2 + \zeta_1^4\zeta_2^2),\end{aligned}$$

which implies, by recalling (A8) and (A14), that

$$\mathrm{Cov}\big(\hat{p}(x),\hat{s}(x)\big) = q(x)\big\{1-q(x)\big\}\sum_{i=1}^{n} V_{n,\ell_1,i}^2 + O(\ell_1^2 + \ell_1\zeta_2^2 + \ell_2\zeta_2^2 + \zeta_1^4\zeta_2^2). \tag{A16}$$

It follows from (A13), (A15), (A16) and the fact $\varsigma = 1 + \zeta_1^2\zeta_2^{-2}$ that

$$\begin{aligned}\mathrm{Var}\big(\hat{\Pi}_{HYB,\varsigma}(x)\big) &= q(x)\big\{1-q(x)\big\}\sum_{i=1}^{n} V_{n,\ell_1,i}^2 \\ &\quad + O(\zeta_1^4\zeta_2^2 + \ell_1^2\zeta_1^2\zeta_2^{-2} + \ell_1\zeta_1^2 + \ell_2\zeta_1^2),\end{aligned} \tag{A17}$$

and from (A8) and (A14) that

$$\begin{aligned}\mathbb{E}\big[\hat{\Pi}_{HYB,\varsigma}(x)\big] &= q(x) + \Big(\frac{1-e^{-\ell_1}}{1-e^{-n\ell_1}}\Big)n\zeta_1^{d+2}\mathcal{J}_1(x;q) + (1-\varsigma)\Big(\frac{1-e^{-\ell_2}}{1-e^{-m\ell_2}}\Big)m\zeta_2^{d+2} \\ &\quad\times\Big[\mathcal{J}_2(x;q) + n\Big(\frac{1-e^{-\ell_1}}{1-e^{-n\ell_1}}\Big)\zeta_1^{d+2}\mathcal{J}_2\big(x;\mathcal{J}_1(\cdot;q)\big)\Big] \\ &\quad + \varsigma\,O(\ell_1\zeta_1^2) + (1-\varsigma)O(\ell_1\zeta_1^2 + \ell_2\zeta_2^2) \\ &= q(x) + \big\{(n\ell_1)^{-2/d} + (1-\varsigma)(m\ell_2)^{-2/d}\big\}\mathcal{V}_d^{-1-2/d}\mathcal{J}_1(x;q) \\ &\quad + O(\ell_1\zeta_1^2 + \ell_2\zeta_1^2 + \zeta_1^2\zeta_2^2).\end{aligned} \tag{A18}$$

Substitution of the expressions (A17), (A18), noting that $\sum_{i=1}^{n} V_{n,\ell_1,i}^2 = 2^{-1}\ell_1\{1+O(\ell_1)\}$ and $q(x) = 1/2$ on $\mathcal{S}$, and invoking asymptotic normality of $\hat{\Pi}_{HYB,\varsigma}(x)$ uniformly over x in a neighbourhood of $\mathcal{S}$, Steps 3 to 5 of the proof of Samworth (2012, Theorem 1) can be adapted to obtain

$$REGRET_{\mathcal{R}}\big(\mathscr{T}^E_{n,m,\varsigma,\boldsymbol{\ell}}\big) = \mathscr{R}^V_{\mathcal{R}}\big(\mathscr{T}^E_{n,m,\varsigma,\boldsymbol{\ell}}\big) + \mathscr{R}^B_{\mathcal{R}}\big(\mathscr{T}^E_{n,m,\varsigma,\boldsymbol{\ell}}\big),$$

where

$$\begin{aligned}\mathscr{R}^V_{\mathcal{R}}\big(\mathscr{T}^E_{n,m,\varsigma,\boldsymbol{\ell}}\big) &= \int_{\mathcal{S}} \|\nabla q(x)\|^{-1} f_\pi(x)\mathrm{Var}\big(\hat{\Pi}_{HYB,\varsigma}(x)\big)\,d\Omega_{d-1}(x)\big\{1+o(1)\big\} \\ &= 8^{-1}\ell_1\{1+o(1)\}\int_{\mathcal{S}} \|\nabla q(x)\|^{-1} f_\pi(x)\,d\Omega_{d-1}(x) + O(\zeta_1^4\zeta_2^2 + \ell_2\zeta_1^2)\end{aligned} \tag{A19}$$

and

$$\begin{aligned}
\mathscr{R}_{\mathcal{R}}^{B}\big(\mathscr{T}_{n,m,\varsigma,\boldsymbol{\ell}}^{E}\big) &= \int_{\mathcal{S}} \|\nabla q(x)\|^{-1} f_{\pi}(x)\big\{\mathbb{E}\big[\hat{\Pi}_{HYB,\varsigma}(x)\big] - q(x)\big\}^{2} d\Omega_{d-1}(x)\big\{1+o(1)\big\} \\
&= d^{-2}\mathcal{V}_{d}^{-4/d}\Gamma(1+2/d)^{2}\{1+o(1)\} \\
&\quad \times \int_{\mathcal{S}} \|\nabla q(x)\|^{-1} f_{\pi}(x)^{-1-4/d}\big\{\nabla q(x)^{\top}\nabla f_{\pi}(x) + f_{\pi}(x)\mathrm{tr}\left(\nabla^{2} q(x)\right)/2\big\}^{2} \\
&\quad \times \big[(n\ell_1)^{-2/d} + (1-\varsigma)(m\ell_2)^{-2/d} + O(\ell_1\zeta_1^2 + \ell_2\zeta_1^2 + \zeta_1^2\zeta_2^2)\big]^{2} d\Omega_{d-1}(x) \\
&= O(\ell_1\zeta_1^2 + \ell_2\zeta_1^2 + \zeta_1^2\zeta_2^2)^2, \qquad \text{(A20)}
\end{aligned}$$

which proves Theorem 1.

A.2 Proof of Corollary 2

We outline a general strategy for minimising θ_n subject to (7). Note that we may write

$$\theta_n = \theta_n(\boldsymbol{\ell}) = \sum_{i\in\mathcal{N}^+} \tilde{\theta}_i(n,m,\ell_1)\ell_2^{\alpha_i} + \sum_{i\in\mathcal{N}^-} \tilde{\theta}_i(n,m,\ell_1)\ell_2^{-\alpha_i} + \sum_{i\in\mathcal{N}^0} \tilde{\theta}_i(n,m,\ell_1),$$

for some constants $\alpha_i > 0$ and some positive functions $\tilde{\theta}_i(n,m,\ell_1)$. Then we have, subject to the bounds (7),

$$\begin{aligned}
\min_{\boldsymbol{\ell}}\{\theta_n\} &\asymp \min_{\ell_1}\Big\{ \sum_{i\in\mathcal{N}^+} \tilde{\theta}_i(n,m,\ell_1) m^{\alpha_i(d\epsilon-1)} + \sum_{i\in\mathcal{N}^-} \tilde{\theta}_i(n,m,\ell_1)\{m^{\alpha_i\epsilon} + m^{\alpha_i}(n\ell_1)^{-\alpha_i}\} \\
&\qquad + \sum_{i\in\mathcal{N}^0} \tilde{\theta}_i(n,m,\ell_1) + \sum_{(i,j)\in\mathcal{N}^+\times\mathcal{N}^-} \big\{\tilde{\theta}_i(n,m,\ell_1)^{\alpha_j}\tilde{\theta}_j(n,m,\ell_1)^{\alpha_i}\big\}^{1/(\alpha_i+\alpha_j)}\Big\} \\
&\asymp \min_{\ell_1}\{\theta_n^{**}(\ell_1)\} \text{ say,}
\end{aligned}$$

where $\theta_n^{**}(\ell_1)$ depends only on (n,m,ℓ_1) and can, analogous to $\theta_n(\boldsymbol{\ell})$, be expressed as

$$\theta_n^{**}(\ell_1) = \sum_{i\in\mathcal{M}^+} \tilde{\psi}_i(n,m)\ell_1^{\beta_i} + \sum_{i\in\mathcal{M}^-} \tilde{\psi}_i(n,m)\ell_1^{-\beta_i} + \sum_{i\in\mathcal{M}^0} \tilde{\psi}_i(n,m),$$

for constants $\beta_i > 0$ and some positive functions $\tilde{\psi}_i(n,m)$. It then follows that

$$\begin{aligned}
\min_{\boldsymbol{\ell}}\{\theta_n\} &\asymp \theta_n^* \\
&\asymp \sum_{i\in\mathcal{M}^+} \tilde{\psi}_i(n,m) n^{\beta_i(d\epsilon-1)} + \sum_{i\in\mathcal{M}^-} \tilde{\psi}_i(n,m) n^{\beta_i\epsilon} \\
&\quad + \sum_{i\in\mathcal{M}^0} \tilde{\psi}_i(n,m) + \sum_{(i,j)\in\mathcal{M}^+\times\mathcal{M}^-} \big\{\tilde{\psi}_i(n,m)^{\beta_j}\tilde{\psi}_j(n,m)^{\beta_i}\big\}^{1/(\beta_i+\beta_j)}.
\end{aligned}$$

The optimal solution ℓ_1^* can be traced by identifying the dominating term in θ_n^*. Specifically, we set $\ell_1^* = n^{d\epsilon-1}$, $n^{-\epsilon}$ and $\big\{\tilde{\psi}_j(n,m)/\tilde{\psi}_i(n,m)\big\}^{1/(\beta_i+\beta_j)}$ if the dominating term arises from the index sets $\mathcal{M}^+$, $\mathcal{M}^-$ and $\mathcal{M}^+\times\mathcal{M}^-$, respectively. With ℓ_1^* thus set, the optimal solution ℓ_2^* can be derived in a similar fashion by identifying the index set, $\mathcal{N}^+$, $\mathcal{N}^-$ or $\mathcal{N}^+\times\mathcal{N}^-$, which contains the dominating term in $\theta_n^{**}(\ell_1^*)$.

A.3 Proof of Theorem 3

The proof follows closely that of Theorem 1. We highlight below the steps which require different treatments. Throughout the proof we denote by C a generic positive constant which may vary from occasion to occasion.

Recall that under the k-nearest neighbour setting, we have

$$\hat{p}(x) = k_1^{-1} \sum_{i=1}^{k_1} \mathbf{1}\{Y_{(i)}(x) = 1\} \quad \text{and} \quad \hat{s}(x) = k_2^{-1} \sum_{j=1}^{k_2} \hat{p}(\boldsymbol{Z}_{(j)}(x)).$$

Define $M_{\alpha,N,k}$ and $\tilde{M}_{N,k} = I_{N,k} + II_{N,k}$ as in (A1) and (A2), respectively, with $V_{N,\ell,i}$ replaced by $k^{-1}\mathbf{1}\{i \le k\}$.

Write $p^*(\cdot|M,p)$ for the binomial (M,p) mass function. Define, for $a > 0$, $x \in \mathbb{R}^d$ and any positive integers $i \le M+1$, $q_{M,i,a}(x) = \sum_{j=0}^{i-1} p^*\big(j|M, p_a(x)\big)$. By Bernstein's inequality,

$$q_{M,i,a}(x) \begin{cases} \ge 1 - e^{-M\{i/M - p_a(x)\}^2/[p_a(x)\{1-p_a(x)\} + \{i/M - p_a(x)\}/3]}, & p_a(x) \le i/M, \\ \le e^{-M\{i/M - p_a(x)\}^2/[p_a(x)\{1-p_a(x)\} + \{p_a(x) - i/M\}/3]}, & p_a(x) \ge i/M. \end{cases} \tag{A21}$$

Let $\zeta_i = \{(N-1)\mathcal{V}_d/i\}^{-1/d}$. For $N^{\eta/4} < i < N^{1-\eta}$, we have

$$\begin{aligned} p_{\zeta_i r}(x) = {} & i(N-1)^{-1} r^d f_\pi(x) + \{i/(N-1)\}^{1+2/d} r^{d+2} (2d+4)^{-1} \mathcal{V}_d^{-2/d} \mathrm{tr}\left(\nabla^2 f_\pi(x)\right) \\ & + O(\zeta_i^{d+4} r^{d+4}). \end{aligned} \tag{A22}$$

It follows from (A21) and (A22) that

$$\begin{aligned} q_{N-1,i,\zeta_i r}(x) &\ge 1 - \exp\Big[- \frac{(N-1)\{i/(N-1) - p_{\zeta_i r}(x)\}^2}{p_{\zeta_i r}(x)\{1 - p_{\zeta_i r}(x)\} + \{i/(N-1) - p_{\zeta_i r}(x)\}/3} \Big] \\ &\ge 1 - \exp\Big[- \frac{i\big\{\{1 - f_\pi(x) r^d\}^2 + O((i/N)^{2/d})\big\}}{1 + O\{(i/N)^{2/d}\}} \Big] \\ &\ge 1 - e^{-Ci(i/N)^{4/d}(\log N)^2} \end{aligned} \tag{A23}$$

for $0 < r < f_\pi(x)^{-1/d}\big\{1 - (i/N)^{2/d} \log N\big\}$, and

$$\begin{aligned} q_{N-1,i,\zeta_i r}(x) &\le \exp\Big[- \frac{(N-1)\{i/(N-1) - p_{\zeta_i r}(x)\}^2}{p_{\zeta_i r}(x)\{1 - p_{\zeta_i r}(x)\} + \{p_{\zeta_i r}(x) - i/(N-1)\}/3} \Big] \\ &\le \exp\Big[- \frac{i\{f_\pi(x) r^d (1 - 2^{-1}(i/N)^{2/d} \log N) - 1\}^2}{(3/2) r^d - 1/3} \Big] \\ &\le e^{-Ci(i/N)^{4/d}(\log N)^2} \end{aligned} \tag{A24}$$

for $r > f_\pi(x)^{-1/d}\big\{1 + (i/N)^{2/d} \log N\big\}$.

Assume henceforth that $(k/N)N^{\epsilon} + N^{4/(d+4)}/k = O(1)$. It follows by normal approximation to the binomial distribution function and the bounds (A23) and (A24) that

$$\begin{aligned} M_{\alpha,N,k}(z;h) &= k^{1-\alpha}h(z) + k^{-\alpha}N\int\{h(z+t)-h(z)\}f_{\pi}(z+t)q_{N-1,k,\|t\|}(z)\,dt \\ &= k^{1-\alpha}h(z) + k^{-\alpha}N\zeta_k^d\Big[\zeta_k^2\mathcal{J}_{N-1,k}(z;h) + O\big\{\zeta_k^4k^{-1/2}\log N \\ &\quad + k^{-1/2}(\log N)^{-1}e^{-Ck\zeta_k^4(\log N)^2} + \zeta_k^2e^{-Ck\zeta_k^4(\log N)^2}\big\}\Big], \end{aligned} \tag{A25}$$

where

$$\begin{aligned} \mathcal{J}_{N-1,k}(z;h) = \zeta_k^{-2}\int &\{h(z+\zeta_ku)-h(z)\}f_{\pi}(z+\zeta_ku) \\ &\times\Phi\Big(\frac{k-1-(N-1)p_{\zeta_k\|u\|}(z)}{\sqrt{(N-1)p_{\zeta_k\|u\|}(z)\{1-p_{\zeta_k\|u\|}(z)\}}}\Big)\,du \asymp 1 \end{aligned}$$

and Φ denotes the standard normal distribution function.

Define, for $\alpha,\beta>0$ and $z_1,z_2\in\mathbb{R}^d$, $\Psi_{N,\alpha,\beta,z_1,z_2}$ to be the joint distribution function of (N_1+N_2, N_1+N_3), where (N_1,N_2,N_3,N_4) has the multinomial mass function $\mathcal{M}_{N,\alpha,\beta,z_1,z_2}$. In the rest of the proof we shall use the following facts related to normal approximation to $\Psi_{N,\alpha,\beta,z_1,z_2}$. For α,β sufficiently small, a multivariate Berry-Esseen bound (Raič, 2019) can be invoked to show that

$$\big|\mathbb{P}\big((N_1+N_2,N_1+N_3)\in\mathcal{S}\big) - \mathbb{P}\big(\boldsymbol{W}\in\mathcal{S}\big)\big| \le CN^{-1/2}\big\{p_{\alpha}(z_1)^{-1/2}+p_{\beta}(z_2)^{-1/2}\big\} \tag{A26}$$

for any measurable convex $\mathcal{S}\subset\mathbb{R}^2$, where $\boldsymbol{W}=(W_1,W_2)$ denotes a bivariate normal vector with the joint distribution function

$$\begin{aligned} \mathbb{P}(W_1\le w_1, W_2\le w_2) &= \Phi\Big(\frac{w_1-Np_{\alpha}(z_1)}{\sqrt{Np_{\alpha}(z_1)\{1-p_{\alpha}(z_1)\}}}\Big)\Phi\Big(\frac{w_2-Np_{\beta}(z_2)}{\sqrt{Np_{\beta}(z_2)\{1-p_{\beta}(z_2)\}}}\Big) \\ &\quad - \varrho_{\alpha,\beta}(z_1,z_2)\,\phi\Big(\frac{w_1-Np_{\alpha}(z_1)}{\sqrt{Np_{\alpha}(z_1)\{1-p_{\alpha}(z_1)\}}}\Big)\phi\Big(\frac{w_2-Np_{\beta}(z_2)}{\sqrt{Np_{\beta}(z_2)\{1-p_{\beta}(z_2)\}}}\Big) \\ &\quad\times\frac{(w_1-Np_{\alpha}(z_1))(w_2-Np_{\beta}(z_2))}{N\sqrt{p_{\alpha}(z_1)\{1-p_{\alpha}(z_1)\}p_{\beta}(z_2)\{1-p_{\beta}(z_2)\}}} + O\big\{p_{\alpha}(z_1)^2+p_{\beta}(z_2)^2\big\} \end{aligned}$$

and ϕ denotes the standard normal density function.

Analgous to (A4) and (A5), we may write

$$\begin{aligned} I_{N,k}(z_1,z_2;H) &= Nk^{-2}\int H_0(z_1+t)f_{\pi}(z_1+t) \\ &\qquad\times\Psi_{N-1,\|t\|,\|z_1-z_2+t\|,z_1,z_2}(k-1,k-1)\,dt, \end{aligned} \tag{A27}$$

$$\begin{aligned} II_{N,k}(z_1,z_2;H) &= N(N-1)k^{-2}\int H(z_1+t,z_2+s)f_{\pi}(z_1+t)f_{\pi}(z_2+s) \\ &\qquad\times\Big\{\Psi_{N-2,\|t\|,\|s\|,z_1,z_2}(k-1,k-1)+Rem_1(\|t\|,\|s\|,z_1,z_2)\Big\}\,dt\,ds, \end{aligned} \tag{A28}$$

where

$$\begin{aligned}
Rem_1(\|t\|,\|s\|,z_1,z_2) &= \mathbf{1}\{\|z_1-z_2+t\|<\|s\|,\ \|z_2-z_1+s\|<\|t\|\} \\
&\quad\times\sum_{i=0}^{k-1}\mathcal{M}_{N-2,\|t\|,\|s\|,z_1,z_2}(k-1-i,i,i,N-k-i-1) \\
&\quad-\mathbf{1}\{\|z_1-z_2+t\|<\|s\|\} \\
&\quad\times\sum_{i\le i'}\mathcal{M}_{N-2,\|t\|,\|s\|,z_1,z_2}(k-1-i',i,i',N-k-i-1) \\
&\quad-\mathbf{1}\{\|z_2-z_1+s\|<\|t\|\} \\
&\quad\times\sum_{i\ge i'}\mathcal{M}_{N-2,\|t\|,\|s\|,z_1,z_2}(k-1-i,i,i',N-k-i'-1).
\end{aligned}$$

We first establish a bound on the last term in (A28). Letting $\mathcal{E}(z_1,z_2)=\big\{(u,v)\in\mathbb{R}^d\times\mathbb{R}^d: f_\pi(z_1)^{1/d}\|u\|\le[1+f_\pi(z_2)\|v\|^d]^{1/d}+(k/N)^{2/d}\log N\big\}$, we have

$$\begin{aligned}
&\sum_{i\le i'}\mathcal{M}_{N-2,\zeta_k\|u\|,\zeta_k\|v\|,z_1,z_2}(k-1-i',i,i',N-k-i-1) \\
&\quad\le\ p^*\big(k-1\big|N-2,p_{\zeta_k\|v\|}(z_2)\big)\sum_{i=0}^{k-1}p^*\left(i\middle|N-k-1,\frac{p_{\zeta_k\|u\|}(z_1)-\varrho_{\zeta_k\|u\|,\zeta_k\|v\|}(z_1,z_2)}{1-p_{\zeta_k\|v\|}(z_2)}\right) \\
&\quad\le\ p^*\big(k-1\big|N-2,p_{\zeta_k\|v\|}(z_2)\big) \\
&\qquad\times\Big[\mathbf{1}\big\{(u,v)\in\mathcal{E}(z_1,z_2)\big\}+\mathbf{1}\big\{(u,v)\notin\mathcal{E}(z_1,z_2)\big\}e^{-\frac{k\{f_\pi(z_1)\|u\|^d(1-2^{-1}(k/N)^{2/d}\log N)-1\}^2}{(3/2)f_\pi(z_1)\|u\|^d-1/3}}\Big].
\end{aligned}$$

It then follows by normal approximation to $p^*\big(k-1\big|N-2,p_{\zeta_k\|v\|}(z_2)\big)$ (Siotani and Fujikoshi, 1984) that for $\alpha,\beta\ge0$,

$$\begin{aligned}
&\iint\|u\|^\alpha\|v\|^\beta\sum_{i\le i'}\mathcal{M}_{N-2,\zeta_k\|u\|,\zeta_k\|v\|,z_1,z_2}(k-1-i',i,i',N-k-i-1)\,du\,dv \\
&\quad\le\ C\int\|v\|^\beta p^*\big(k-1\big|N-2,p_{\zeta_k\|v\|}(z_2)\big)\big\{1+\|v\|^{d+\alpha}+e^{-Ck\zeta_k^4(\log N)^2}\big\}\,dv \\
&\quad\le\ Ck^{-1/2}\int_{|f_\pi(z_2)^{1/d}\|v\|-1|\le k^{-1/2}(\log N)^{2/3}}\|v\|^\beta\big(1+\|v\|^{d+\alpha}\big)\phi\Big(\frac{k^{1/2}\{1-f_\pi(z_2)\|v\|^d\}}{f_\pi(z_2)^{1/2}\|v\|^{d/2}}\Big) \\
&\qquad\times\big\{1+O(k^{-1/2}(\log N)^2)\big\}\,dv+Ce^{-C(\log N)^{4/3}} \\
&\quad\le\ Ck^{-1}(\log N)^{2/3}. \qquad\text{(A29)}
\end{aligned}$$

Similarly, the same bound can be shown to hold if the summation in the integrand in (A29) is taken over $i\ge i'$ or $0\le i\le k-1$, respectively.

Consider next the distribution function $\Psi_{N,\alpha,\beta,z_1,z_2}$. Writing $\varepsilon^* = (k/N)^{2/d}\log N$ and using (A23), (A24) and (A26), we have

$$\begin{aligned}
&\Psi_{N,\zeta_k r_1,\zeta_k r_2,z_1,z_2}(k-1,k-1)\\
&\quad= \mathbf{1}\big\{|f_\pi(z_1)^{1/d}r_1 - 1| < \varepsilon^*, f_\pi(z_2)^{1/d}r_2 < 1-\varepsilon^*\big\}\\
&\qquad\times\Big[\Phi\Big(\frac{k-1-Np_{\zeta_k r_1}(z_1)}{\sqrt{Np_{\zeta_k r_1}(z_1)\{1-p_{\zeta_k r_1}(z_1)\}}}\Big) + O\big\{\zeta_k^d + k^{-1/2} + e^{-Ck\zeta_k^4(\log N)^2}\big\}\Big]\\
&\qquad+\mathbf{1}\big\{|f_\pi(z_2)^{1/d}r_2 - 1| < \varepsilon^*, f_\pi(z_1)^{1/d}r_1 < 1-\varepsilon^*\big\}\\
&\qquad\times\Big[\Phi\Big(\frac{k-1-Np_{\zeta_k r_2}(z_2)}{\sqrt{Np_{\zeta_k r_2}(z_2)\{1-p_{\zeta_k r_2}(z_2)\}}}\Big) + O\big\{\zeta_k^d + k^{-1/2} + e^{-Ck\zeta_k^4(\log N)^2}\big\}\Big]\\
&\qquad+\mathbf{1}\big\{|f_\pi(z_j)^{1/d}r_j - 1| < \varepsilon^*,\ j=1,2\big\}\\
&\qquad\times\Big[\Phi\Big(\frac{k-1-Np_{\zeta_k r_1}(z_1)}{\sqrt{Np_{\zeta_k r_1}(z_1)\{1-p_{\zeta_k r_1}(z_1)\}}}\Big)\Phi\Big(\frac{k-1-Np_{\zeta_k r_2}(z_2)}{\sqrt{Np_{\zeta_k r_2}(z_2)\{1-p_{\zeta_k r_2}(z_2)\}}}\Big)\\
&\qquad\quad+O(\zeta_k^d + k^{-1/2})\Big] + \mathbf{1}\big\{\max_{j\in\{1,2\}}\{f_\pi(z_j)^{1/d}r_j\} > 1+\varepsilon^*\big\}O\big\{e^{-Ck\zeta_k^4(\log N)^2}\big\}\\
&\qquad+\mathbf{1}\big\{\max_{j\in\{1,2\}}\{f_\pi(z_j)^{1/d}r_j\} < 1-\varepsilon^*\big\}\Big[1+O\big\{e^{-Ck\zeta_k^4(\log N)^2}\big\}\Big]. \qquad \text{(A30)}
\end{aligned}$$

Define also $\mathcal{K}_{N-1,k}(z) = \int f_\pi(z+\zeta_k u)\Phi\Big(\frac{k-1-(N-1)p_{\zeta_k\|u\|}(z)}{\sqrt{(N-1)p_{\zeta_k\|u\|}(z)\{1-p_{\zeta_k\|u\|}(z)\}}}\Big)\,du$.

We are now ready to establish expansions for the mean and variance of $\hat{s}(x)$. Let $\xi_1 = (n\mathcal{V}_d/k_1)^{-1/d}$ and $\xi_2 = (m\mathcal{V}_d/k_2)^{-1/d}$, so that $k_j^{-1} = O(\xi_j^4)$, $j=1,2$, according to (9). It follows by setting $\alpha = 1$ in (A25) that, uniformly over $x \in \mathcal{X}_{n,\eta}$,

$$\begin{aligned}
\mathbb{E}\big[\hat{s}(x)\big] &= M_{1,m,k_2}\big(x; M_{1,n,k_1}(\cdot;q)\big)\\
&= q(x) + k_1^{-1}n\xi_1^{d+2}\mathcal{J}_{n-1,k_1}(x;q) + k_2^{-1}m\xi_2^{d+2}\mathcal{J}_{m-1,k_2}(x;q)\\
&\quad + k_1^{-1}k_2^{-1}nm\xi_1^{d+2}\xi_2^{d+2}\mathcal{J}_{m-1,k_2}\big(x;\mathcal{J}_{n-1,k_1}(\cdot;q)\big)\\
&\quad + O\big(k_1^{-1/2}\xi_1^4\log n + k_2^{-1/2}\xi_2^4\log m\big). \qquad \text{(A31)}
\end{aligned}$$

Consider next $\mathbb{E}\big[\hat{s}(x)^2\big] = \tilde{M}_{m,k_2}\big(x,x;\tilde{M}_{n,k_1}(\cdot,\cdot;q^{\otimes})\big)$, which has, by (A27) and (A28), the expansion

$$\begin{aligned}
&k_2^{-1}q(x)^2 + m(m-1)k_2^{-2}\xi_2^{2d}\\
&\quad\times\iint f_\pi(x+\xi_2 u)f_\pi(x+\xi_2 v)\{\tilde{M}_{n,k_1}(x+\xi_2 u, x+\xi_2 v; q^{\otimes}) - \tilde{M}_{n,k_1}(x,x;q^{\otimes})\}\\
&\qquad\times\Big\{\Psi_{m-2,\xi_2\|u\|,\xi_2\|v\|,x,x}(k_2-1,k_2-1) + Rem_1(\xi_2\|u\|,\xi_2\|v\|,x,x)\Big\}\,du\,dv\\
&\quad+\tilde{M}_{n,k_1}(x,x;q^{\otimes})\Big\{1 - mk_2^{-2}\xi_2^d\int f_\pi(x+\xi_2 u)\Psi_{m-1,\xi_2\|u\|,\xi_2\|u\|,x,x}(k_2-1,k_2-1)\,du\Big\}\\
&\qquad\qquad + O(k_2^{-1}\xi_2^2 + k_1^{-1}k_2^{-1}). \qquad \text{(A32)}
\end{aligned}$$

Applying the results (A27)–(A30), we obtain

$$\begin{aligned}
&\tilde{M}_{n,k_1}(x+\xi_2 u, x+\xi_2 v; q^{\otimes}) \\
&= nk_1^{-2}\xi_1^d \int \{q(x+\xi_2 u+\xi_1 w) - q(x+\xi_2 u)\} f_\pi(x+\xi_2 u+\xi_1 w) \\
&\quad \times \Psi_{n-1,\xi_1\|w\|,\|\xi_2(u-v)+\xi_1 w\|,x+\xi_2 u,x+\xi_2 v}(k_1-1,k_1-1)\,dw \\
&\quad + k_1^{-2} n(n-1)\xi_1^{2d+2}\big\{\xi_1^2 \mathcal{J}_{n-2,k_1}(x+\xi_2 u;q)\mathcal{J}_{n-2,k_1}(x+\xi_2 v;q) \\
&\quad + q(x+\xi_2 v)\mathcal{J}_{n-2,k_1}(x+\xi_2 u;q)\mathcal{K}_{n-2,k_1}(x+\xi_2 v) \\
&\quad + q(x+\xi_2 u)\mathcal{J}_{n-2,k_1}(x+\xi_2 v;q)\mathcal{K}_{n-2,k_1}(x+\xi_2 u)\big\} + q(x+\xi_2 u)q(x+\xi_2 v) \\
&\quad + k_1^{-2} q(x+\xi_2 u)\big\{1-q(x+\xi_2 v)\big\} \sum_{i,i'=1}^{k_1} \int g^I_{n,i,i',x+\xi_2 u,x+\xi_2 v}(t)\,dt \\
&\quad + O\big\{\xi_1^4(\xi_1^d + k_1^{-1/2})\log n + k_1^{-1}\xi_1^2(\log n)^{2/3}\big\} \qquad \text{(A33)}
\end{aligned}$$

and

$$\begin{aligned}
&\tilde{M}_{n,k_1}(x,x;q^{\otimes}) \\
&= q(x)^2 + k_1^{-1}q(x)\big\{1-q(x)\big\} + O\big\{\xi_1^4(\xi_1^d + k_1^{-1/2})\log n + k_1^{-1}\xi_1^2(\log n)^{2/3}\big\} \\
&\quad + k_1^{-2} n(n-1)\xi_1^{2d+2}\big\{\xi_1^2 \mathcal{J}_{n-2,k_1}(x;q)^2 + 2q(x)\mathcal{J}_{n-2,k_1}(x;q)\mathcal{K}_{n-2,k_1}(x)\big\}. \qquad \text{(A34)}
\end{aligned}$$

Substituting (A33) and (A34) into (A32), using (A29), (A30) and the fact that $\mathcal{K}_{N-1,k}(z) = (k/N)\zeta_k^{-d} + O(k^{-1/2}\zeta_k^2 \log N)$, we have

$$\begin{aligned}
\mathbb{E}\big[\hat{s}(x)^2\big] &= q(x)^2 + k_1^{-1}q(x)\big\{1-q(x)\big\} + 2nmk_1^{-1}k_2^{-1}\xi_1^{d+2}\xi_2^{d+2} \\
&\quad \times \big\{q(x)\mathcal{J}_{m-2,k_2}\big(x;\mathcal{J}_{n-2,k_1}(\cdot;q)\big) + \mathcal{J}_{n-2,k_1}(x;q)\mathcal{J}_{m-2,k_2}\big(x;q\big)\big\} \\
&\quad + n^2k_1^{-2}\xi_1^{2d+2}\big\{\xi_1^2\mathcal{J}_{n-2,k_1}(x;q)^2 + 2k_1 n^{-1}\xi_1^{-d}q(x)\mathcal{J}_{n-2,k_1}(x;q)\big\} \\
&\quad + m^2k_2^{-2}\xi_2^{2d+2}\big\{\xi_2^2\mathcal{J}_{m-2,k_2}(x;q)^2 + 2k_2 m^{-1}\xi_2^{-d}q(x)\mathcal{J}_{m-2,k_2}(x;q)\big\} \\
&\quad + O\big\{k_1^{-1/2}\xi_1^4\log n + \xi_2^4(\xi_2^d + k_2^{-1/2})\log n + \xi_1^2\xi_2^4\big\}. \qquad \text{(A35)}
\end{aligned}$$

Subtracting the square of (A31) from (A35) yields

$$\begin{aligned}
\mathrm{Var}\big(\hat{s}(x)\big) &= k_1^{-1}q(x)\big\{1-q(x)\big\} \\
&\quad + O\big\{k_1^{-1/2}\xi_1^4\log n + \xi_2^4(\xi_2^d + k_2^{-1/2})\log n + \xi_1^2\xi_2^4\big\}, \qquad \text{(A36)}
\end{aligned}$$

uniformly over $x \in \mathcal{X}_{n,\eta}$. Similar, but simpler, arguments show that

$$\begin{aligned}
\mathbb{E}\big[\hat{p}(x)\big] &= M_{1,n,k_1}(x;q) = q(x) + k_1^{-1}n\xi_1^{d+2}\mathcal{J}_{n-1,k_1}(x;q) + O\big(k_1^{-1/2}\xi_1^4\log n\big), \qquad \text{(A37)} \\
\mathrm{Var}\big(\hat{p}(x)\big) &= \tilde{M}_{n,k_1}(x,x;q^{\otimes}) - M_{1,n,k_1}(x;q)^2 \\
&= k_1^{-1}q(x)\big\{1-q(x)\big\} + O\big\{\xi_1^4(\xi_1^d + k_1^{-1/2})\log n\big\}. \qquad \text{(A38)}
\end{aligned}$$

Applying again (A25) with $\alpha = 1$, (A27), (A28), (A33) and (A34), we have

$$\begin{aligned}\mathbb{E}\big[\hat{p}(x)\hat{s}(x)\big] &= M_{1,m,k_2}\big(x; \tilde{M}_{n,k_1}(x,\cdot;q^{\otimes})\big)\\ &= q(x)^2 + k_1^{-1}q(x)\big\{1-q(x)\big\} + 2k_1^{-1}n\xi_1^{d+2}q(x)\mathcal{J}_{n-2,k_1}(x;q)\\ &\quad + k_2^{-1}m\xi_2^{d+2}q(x)\mathcal{J}_{m-1,k_2}(x;q) + k_1^{-2}n^2\xi_1^{2d+4}\mathcal{J}_{n-2,k_1}(x;q)^2 + k_1^{-1}k_2^{-1}nm\xi_1^{d+2}\xi_2^{d+2}\\ &\quad \times\big\{\mathcal{J}_{n-2,k_1}(x;q)\mathcal{J}_{m-1,k_2}\big(x;q\big) + q(x)\mathcal{J}_{m-1,k_2}\big(x;\mathcal{J}_{n-2,k_1}(\cdot;q)\big)\big\}\\ &\quad + O\big\{\xi_2^4k_2^{-1/2}\log m + \xi_1^4(\xi_1^d + k_1^{-1/2})\log n + \xi_1^4\xi_2^2\big\},\end{aligned}$$

so that

$$\begin{aligned}\mathrm{Cov}\big(\hat{p}(x),\hat{s}(x)\big) &= k_1^{-1}q(x)\big\{1-q(x)\big\}\\ &\quad + O\big\{k_2^{-1/2}\xi_2^4\log n + \xi_1^4(\xi_1^d + k_1^{-1/2})\log n + \xi_1^4\xi_2^2\big\}.\end{aligned} \tag{A39}$$

Combining the results (A31), (A36), (A37), (A38), (A39) and noting that $\varsigma = 1 + \xi_1^2\xi_2^{-2}$, we obtain

$$\begin{aligned}\mathrm{Var}\big(\hat{\Pi}_{HYB,\varsigma}(x)\big) &= k_1^{-1}q(x)\big\{1-q(x)\big\}\\ &\quad + O\big\{\xi_1^6 + \xi_1^4(\xi_2^d + k_1^{-1/2})\log n + k_2^{-1/2}\xi_1^2\xi_2^2\log n\big\}\end{aligned} \tag{A40}$$

and, by substituting the leading terms of the functions $\mathcal{J}_{n-1,k_1}$ and $\mathcal{J}_{m-1,k_2}$,

$$\begin{aligned}\mathbb{E}\big[\hat{\Pi}_{HYB,\varsigma}(x)\big] &= q(x) + \big\{(k_1/n)^{2/d} + (1-\varsigma)(k_2/m)^{2/d}\big\}(d+2)^{-1}\mathcal{V}_d^{-2/d}\\ &\quad \times f_\pi(x)^{-1-2/d}\big\{f_\pi(x)\mathrm{tr}\,(\nabla^2 q(x))/2 + \nabla f_\pi(x)^\top\nabla q(x)\big\} + O(\xi_1^2\xi_2^2).\end{aligned} \tag{A41}$$

Theorem 3 then follows by applying the same arguments as given in the last part of the proof of Theorem 1, with (A17) and (A18) replaced by (A40) and (A41), respectively.

A.4 Proof of Corollary 4

Corollary 4 is proved by applying the same strategy as has been outlined in the proof of Corollary 2. We omit the details here.

A.5 Supplement to Corollary 2

Under the conditions of Corollary 2, the optimal rates of (ℓ_1, ℓ_2) which minimise θ_n are given below:

(i) If $m \preceq n^{2/\{d+4-(d^2+6d+4)\epsilon\}}$, then

$$\ell_1 \asymp n^{-2/(d+2)}m^{-d(1-d\epsilon)/(d+2)}, \quad \ell_2 \asymp m^{-1+d\epsilon}.$$

(ii) If $n^{2/\{d+4-(d^2+6d+4)\epsilon\}} \preceq m \preceq n^{(d+4)/(d+6)}$, then

$$\ell_1 \asymp \big(n^{4d+4}m^{2d}\big)^{-1/(d^2+6d+4)}, \quad \ell_2 \asymp \big(n^{2d}m^{2d+4}\big)^{-1/(d^2+6d+4)}.$$

(iii) If $m \succeq n^{(d+4)/(d+6)}$, then

$$\ell_1 \asymp n^{-6/(d+6)}, \quad \ell_2 \asymp n^{d/(d+6)}m^{-1}.$$

The change of the optimal $\boldsymbol{\ell}$ with the unlabelled sample size m is plotted in Figure 9 on the $\log_n$ scale.

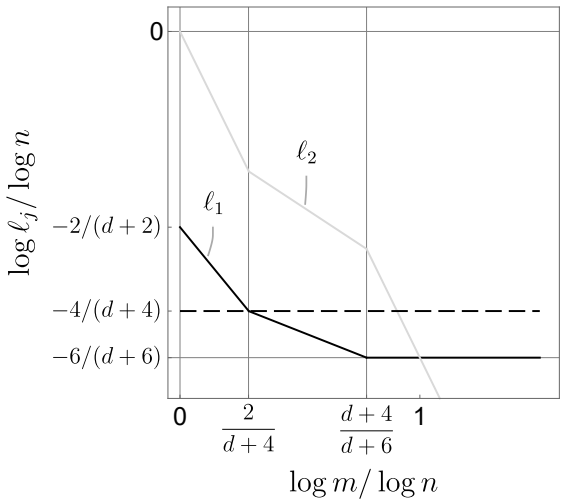

Figure 9: Plots of optimal $\log \ell_j / \log n$ against $\log m / \log n$, with ϵ set at 0.001. Dashed line indicates level $-4/(d+4)$ required by optimally weighted nearest neighbour classifier trained on $\mathcal{L}_S$ alone.

A.6 Supplement to Corollary 4

Under the conditions of Corollary 4, the optimal rates of (k_1, k_2) which minimise ψ_n are given below:

(i) For $d \geq 5$,

(i.1) if $m \preceq \left(n^{2/(d+4)} \log n\right)^{2/\{1-\epsilon(1-4/d)\}}$, then $k_1 \asymp n^{4/(d+4)}$, $k_2 \asymp m^{1-\epsilon}$;

(i.2) if $(n^{2/(d+4)} \log n)^{2/\{1-\epsilon(1-4/d)\}} \preceq m \preceq (n^{2/(d+4)} \log n)^{2/\{1-\epsilon(d+8)/(d+4)\}}$, then

$$k_1 \asymp \{n^4 m^{d-\epsilon(d-4)} (\log n)^{-2d}\}^{1/(2d+4)}, \quad k_2 \asymp m^{1-\epsilon};$$

(i.3) if $(n^{2/(d+4)} \log n)^{2/\{1-\epsilon(d+8)/(d+4)\}} \preceq m \preceq n^{(d+12)/(2d+12)} (\log n)^2$, then

$$k_1 \asymp \{nm(\log n)^{-2}\}^{4/(d+8)}, \quad k_2 \asymp \{nm(\log n)^{2+d/2}\}^{4/(d+8)};$$

(i.4) if $m \succeq n^{(d+12)/(2d+12)} (\log n)^2$, then $k_1 \asymp n^{6/(d+6)}$, $k_2 \asymp n^{-d/(2d+12)} m$.

(ii) For $d = 3, 4$,

(ii.1) if $m \preceq n^{1/2} (\log n)^{1+d/4}$, then $k_1 \asymp n^{4/(d+4)}$, $k_2 \asymp m^{4/(d+4)}$;

(ii.2) if $n^{1/2} (\log n)^{1+d/4} \preceq m \preceq (n \log n)^{(d+4)/(d+6)}$, then

$$k_1 \asymp \{n^2 m^{4d/(d+4)} (\log n)^{-d}\}^{1/(d+2)}, \quad k_2 \asymp m^{4/(d+4)};$$

(ii.3) if $(n \log n)^{(d+4)/(d+6)} \preceq m \preceq n^{(d+4)/(d+6)} (\log n)^2$, then

$$k_1 \asymp \{n^{8-d} m^d (\log n)^{-2d}\}^{1/(d+8)}, \quad k_2 \asymp \{(n \log n)^d m^{-d-4}\}^{-2/(d+8)};$$

(ii.4) if $m \succeq n^{(d+4)/(d+6)}(\log n)^2$, then $k_1 \asymp n^{6/(d+6)}$, $k_2 \asymp m^{4/(d+4)}$.

(iii) For $d = 1, 2$,

(iii.1) if $m \preceq n^{1/2}(\log n)^{1+d/4}$, then $k_1 \asymp n^{4/(d+4)}$, $k_2 \asymp m^{4/(d+4)}$;

(iii.2) if $n^{1/2}(\log n)^{1+d/4} \preceq m \preceq (n \log n)^{2(d+4)/(16+2d-d^2)}$, then

$$k_1 \asymp \{n^2 m^{4d/(d+4)}(\log n)^{-d}\}^{1/(d+2)}, \quad k_2 \asymp m^{4/(d+4)};$$

(iii.3) if $(n \log n)^{2(d+4)/(16+2d-d^2)} \preceq m \preceq (n \log n)^{(d+4)/(2d+4)}$, then

$$k_1 \asymp \{n^4 m^{d^2/(d+4)}(\log n)^{-d}\}^{1/(d+4)}, \quad k_2 \asymp m^{4/(d+4)};$$

(iii.4) if $m \succeq (n \log n)^{(d+4)/(2d+4)}$, then

$$k_1 \asymp \{n^{d+4}(\log n)^{-d}\}^{1/(2d+4)}, \quad k_2 \asymp (n \log n)^{-d/(2d+4)} m.$$

The above optimal orders of $\boldsymbol{k}$ are displayed in Figure 10.

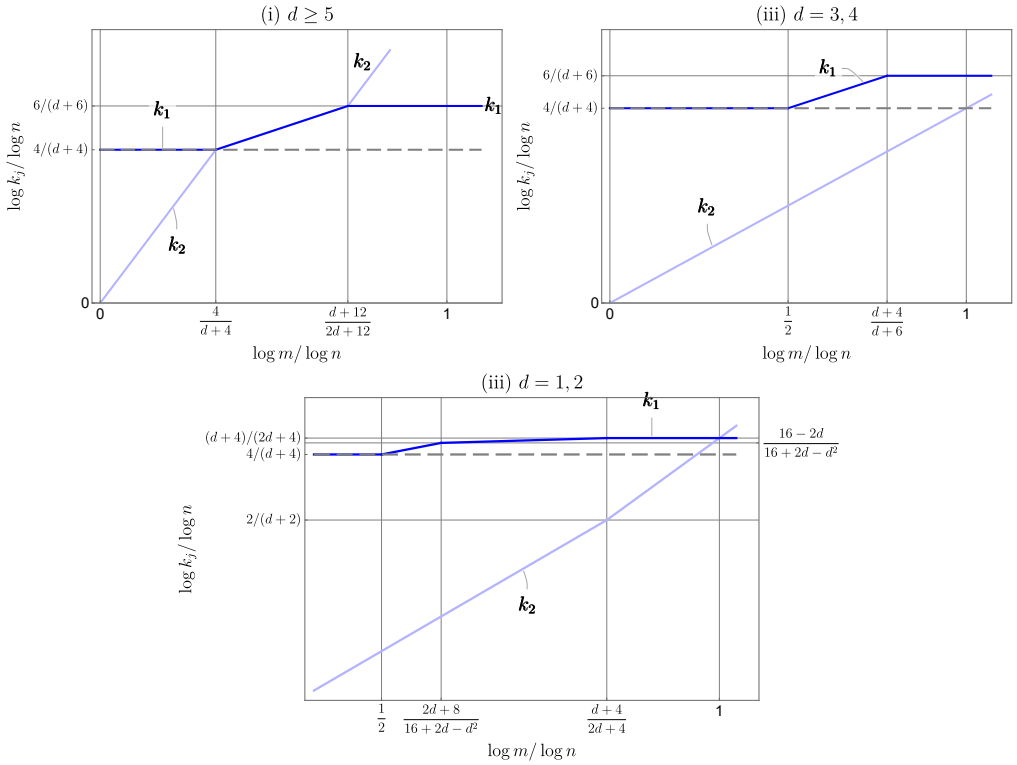

Figure 10: Plots of optimal $\log k_j / \log n$ against $\log m / \log n$, with ϵ set at 0.001. Dashed line indicates level $4/(d+4)$ required by optimal k nearest neighbour classifier trained on $\mathcal{L}_S$ alone.

A.7 Supplement to Corollary 5

Under the conditions of Corollary 5, the optimal rates of (ℓ_1, ℓ_2) which minimise ϑ_n are given below:

(i) For $d \geq 2$,

(i.1) If $m \preceq n^{(d+4)/(d+6)}$, then $\ell_1 \asymp n^{-4/(d+4)} m^{-2d/(d+4)^2}$, $\ell_2 \asymp m^{-4/(d+4)}$.

(i.2) If $m \succeq n^{(d+4)/(d+6)}$, then $\ell_1 \asymp n^{-6/(d+6)}$, $\ell_2 \asymp n^{-2d/\{(d+2)(d+6)\}} m^{-2/(d+2)}$.

(ii) For $d = 1$,

(ii.1) If $m \preceq n^{5/12}$, then $\ell_1 \asymp n^{-4/5} m^{-2/25}$, $\ell_2 \asymp m^{-4/5}$.

(ii.2) If $m \succeq n^{5/12}$, then $\ell_1 \preceq n^{-4/5} m^{(2-3\epsilon)/5}$, $\ell_2 \asymp m^{-\epsilon}$.

A.8 Supplement to Corollary 6

Under the conditions of Corollary 6, the optimal rates of (k_1, k_2) which minimise φ_n are given below:

(i) For $d \geq 8$,

(i.1) If $m \preceq n(\log n)^{-d/2}$, then

$$k_1 \asymp n^{4/(d+4)} \{m(\log n)^{-2}\}^{4d/\{(d+4)(d+8)\}}, \quad k_2 \asymp \{m^4 (\log n)^d\}^{2/(d+8)}.$$

(i.2) If $n(\log n)^{-d/2} \preceq m \preceq n$, then $k_1 \asymp k_2 \asymp \{m^4 (\log n)^d\}^{2/(d+8)}$.

(i.3) If $n \preceq m \preceq \{n^4 (\log n)^d\}^{(d+4)/(2d+16)}$, then $k_1 \asymp k_2 \asymp \{n^4 (\log n)^d\}^{2/(d+8)}$.

(i.4) If $m \succeq \{n^4 (\log n)^d\}^{(d+4)/(2d+16)}$, then

$$k_1 \asymp \{n^4 (\log n)^d\}^{2/(d+8)}, \quad k_2 \asymp m^{4/(d+4)}.$$

(ii) For $5 \leq d \leq 7$,

(ii.1) If $m \preceq n^{(2d+16)/(d+24)} (\log n)^{(d^2+4d+32)/(2d+48)}$, then

$$k_1 \asymp n^{4/(d+4)} \{m(\log n)^{-2}\}^{4d/\{(d+4)(d+8)\}}, \quad k_2 \asymp \{m^4 (\log n)^d\}^{2/(d+8)}.$$

(ii.2) If $n^{(2d+16)/(d+24)} (\log n)^{(d^2+4d+32)/(2d+48)} \preceq m \preceq \{n^4 (\log n)^d\}^{(d+4)/(2d+16)}$, then

$$\{n^4 (\log n)^{-d}\}^{1/(d+4)} m^{2d/(d+4)^2} \preceq k_1 \preceq n m^{-d/(2d+8)} (\log n)^{d/4}, \quad k_2 \asymp m^{4/(d+4)}.$$

(ii.3) If $m \succeq \{n^4 (\log n)^d\}^{(d+4)/(2d+16)}$, then

$$k_1 \asymp \{n^4 (\log n)^d\}^{2/(d+8)}, \quad k_2 \asymp m^{4/(d+4)}.$$

(iii) For $d = 4$,

(iii.1) If $m \preceq (n \log n)^{6/7}$, then $k_1 \asymp (n/\log n)^{1/2} m^{1/6}$, $k_2 \asymp m^{2/3}$.

(iii.2) If $(n \log n)^{6/7} \preceq m \preceq n^{4/3}$, then

$$(n/\log n)^{1/2} m^{1/8} \preceq k_1 \preceq n m^{-1/4}, \quad k_2 \asymp m^{1/2}.$$

(iii.3) If $m \succeq n^{4/3}$, then $k_1 \asymp n^{2/3}$, $k_2 \asymp m^{1/2}$.

(iv) For $2 \leq d \leq 3$,

(iv.1) If $m \preceq (n \log n)^{12/(18-d)}$, then $k_1 \asymp \{n^4 m^{d/3} (\log n)^{-d}\}^{1/(d+4)}$, $k_2 \asymp m^{2/3}$.

(iv.2) If $(n \log n)^{12/(18-d)} \preceq m \preceq (n \log n)^{(d^2+4d)/(4d+8)}$, then

$$\{n^4 (\log n)^{-d}\}^{1/(d+4)} m^{2d/(d+4)^2} \preceq k_1 \preceq n m^{-2/(d+4)}, \quad k_2 \asymp m^{4/(d+4)}.$$

(iv.3) If $m \succeq (n \log n)^{(d^2+4d)/(4d+8)}$, then

$$k_1 \asymp \{n^{d+4} (\log n)^{-d}\}^{1/(2d+4)}, \quad k_2 \asymp m^{4/(d+4)}.$$

(v) For $d = 1$,

(v.1) If $m \preceq (n \log n)^{5/6}$, then $k_1 \asymp n^{4/5} m^{1/25} (\log n)^{-1/5}$, $k_2 \asymp m^{4/5}$.

(v.2) If $m \succeq (n \log n)^{5/6}$, then $k_1 \asymp n^{5/6} (\log n)^{-1/6}$, $k_2 \asymp m^{4/5}$.

References

Raič, M. (2019). A multivariate Berry–Esseen theorem with explicit constants. *Bernoulli*, **25**, 2824–2853.

Siotani, M. and Fujikoshi, Y. (1984). Asymptotic approximations for the distributions of multinomial goodness-of-fit statistics. *Hiroshima Math. J.*, **14**, 115–124.