Biometrika, *pp*. 1–19
Printed in Great Britain

Classification with imperfect training labels

BY TIMOTHY I. CANNINGS
School of Mathematics, University of Edinburgh,
James Clerk Maxwell Building, Edinburgh EH9 3FD, U.K.
timothy.cannings@ed.ac.uk

YINGYING FAN
Department of Data Science and Operations, University of Southern California,
Los Angeles, California 90089, U.S.A.
fanyingy@marshall.usc.edu

AND RICHARD J. SAMWORTH
Statistical Laboratory, University of Cambridge,
Centre for Mathematical Sciences, Cambridge CB3 0WB, U.K.
r.samworth@statslab.cam.ac.uk

SUMMARY

We study the effect of imperfect training data labels on the performance of classification methods. In a general setting, where the probability that an observation in the training dataset is mislabelled may depend on both the feature vector and the true label, we bound the excess risk of an arbitrary classifier trained with imperfect labels in terms of its excess risk for predicting a noisy label. This reveals conditions under which a classifier trained with imperfect labels remains consistent for classifying uncorrupted test data points. Furthermore, under stronger conditions, we derive detailed asymptotic properties for the popular k-nearest neighbour ($k$nn), support vector machine (SVM) and linear discriminant analysis (LDA) classifiers. One consequence of these results is that the $k$nn and SVM classifiers are robust to imperfect training labels, in the sense that the rate of convergence of the excess risks of these classifiers remains unchanged; in fact, our theoretical and empirical results even show that in some cases, imperfect labels may improve the performance of these methods. On the other hand, the LDA classifier is shown to be typically inconsistent in the presence of label noise unless the prior probabilities of each class are equal. Our theoretical results are supported by a simulation study.

Some key words: Label noise; Linear discriminant analysis; Misclassification error; Nearest neighbours; Statistical learning; Support vector machines.

1. INTRODUCTION

Supervised classification is one of the fundamental problems in statistical learning. In the basic, binary setting, the task is to assign an observation to one of two classes, based on a number of previous training observations from each class. Modern applications include, among many others, diagnosing a disease using genomics data (Wright et al., 2015), determining a user's action from smartphone telemetry data (Lara & Labrador, 2013), and detecting fraud based on historical financial transactions (Bolton & Hand, 2002).

In a classification problem it is often the case that the class labels in the training data set are inaccurate. For instance, an error could simply arise due to a coding mistake when the data were recorded. In other circumstances, such as the disease diagnosis application mentioned above, errors may occur due to the fact that, even to an expert, the true labels are hard to determine, especially if there is insufficient information available. Moreover, in modern Big Data applications with huge training data sets, it may be impractical and expensive to determine the true class labels, and as a result the training data labels are often assigned by an imperfect algorithm. Services such as the Amazon Mechanical Turk, e.g. `https://www.mturk.com`, allow practitioners to obtain training data labels relatively cheaply via crowdsourcing. Of course, even after aggregating a large crowd of workers' labels, the result may still be inaccurate. Chen et al. (2015) and Zhang et al. (2016) discuss crowdsourcing in more detail, and investigate strategies for obtaining the most accurate labels given a cost constraint.

The problem of label noise was first studied by Lachenbruch (1966), who investigated the effect of imperfect labels in two-class linear discriminant analysis. Other early works of note include Lachenbruch (1974), Angluin & Laird (1988) and Lugosi (1992).

Frénay & Kabán (2014) and Frénay & Verleysen (2014) provide recent overviews of work on the topic. In the simplest, homogeneous setting, each observation in the training dataset is mislabelled independently with some fixed probability. van Rooyen et al. (2015) study the effects of homogeneous label errors on the performance of empirical risk minimization classifiers, while Long & Servedio (2010) consider boosting methods in this same homogeneous noise setting. Other recent works focus on class-dependent label noise, where the probability that a training observation is mislabelled depends on the true class label of that observation; see Stempfel & Ralaivola (2009), Natarajan et al. (2013), Scott et al. (2013), Blanchard et al. (2016), Liu & Tao (2016) and Patrini et al. (2016). An alternative model assumes the noise rate depends on the feature vector of the observation. Manwani & Sastry (2013) and Ghosh et al. (2015) investigate the properties of empirical risk minimization classifiers in this setting; see also Awasthi et al. (2015). Menon et al. (2016) propose a generalized boundary consistent label noise model, where observations near the optimal decision boundary are more likely to be mislabelled, and study the effects on the properties of the receiver operator characteristics curve.

In the more general setting, where the probability of mislabelling is both feature- and class-dependent, Bootkrajang & Kabán (2012, 2014) and Bootkrajang (2016) study the effect of label noise on logistic regression classifiers, while Li et al. (2017), Patrini et al. (2017) and Rolnick et al. (2017) consider neural network classifiers. On the other hand, Cheng et al. (2017) investigate the performance of an empirical risk minimization classifier in the feature- and class-dependent noise setting when the true class conditional distributions have disjoint support.

Our first goal in the present paper is to provide general theory to characterize the effect of feature- and class-dependent heterogeneous label noise for an arbitrary classifier. We first specify general conditions under which the optimal prediction of a true label and a noisy label are the same for every feature vector. Then, under slightly stronger conditions, we relate the misclassification error when predicting a true label to the corresponding error when predicting a noisy label. More precisely, we show that the excess risk, i.e. the difference between the error rate of the classifier and that of the optimal, Bayes classifier, is bounded above by the excess risk associated with predicting a noisy label multiplied by a constant factor that does not depend on the classifier used; see Theorem 1. Our results therefore provide conditions under which a classifier trained with imperfect labels remains consistent for classifying uncorrupted test data points.

As applications of these ideas, we consider three popular approaches to classification problems, namely the k-nearest neighbour ($k$nn), support vector machine (SVM) and linear discriminant analysis (LDA) classifiers. In the perfectly labelled setting, the $k$nn classifier is consistent

for any data generating distribution and the SVM classifier is consistent when the distribution of the feature vectors is compactly supported. Since the label noise does not change the marginal feature distribution, it follows from our results mentioned in the previous paragraph that these two methods are still consistent when trained with imperfect labels that satisfy our assumptions, which, in the homogeneous noise case, even allow up to 1/2 of the training data to be labelled incorrectly. On the other hand, for the LDA classifier with Gaussian class-conditional distributions, we derive the asymptotic risk in the homogeneous label noise case. This enables us to deduce that the LDA classifier is typically not consistent when trained with imperfect labels, unless the class prior probabilities are equal to 1/2.

Our second main contribution is to provide greater detail on the asymptotic performance of the $k$nn and SVM classifiers in the presence of label noise, under stronger conditions on the data generating mechanism and noise model. In particular, for the $k$nn classifier, we derive the asymptotic limit for the ratio of the excess risks of the classifier trained with imperfect and perfect labels, respectively. This reveals the nice surprise that using imperfectly-labelled training data can in fact improve the performance of the $k$nn classifier in certain circumstances. To the best of our knowledge, this is the first formal result showing that label noise can help with classification. For the SVM classifier, we provide conditions under which the rate of convergence of the excess risk is unaffected by label noise, and show empirically that this method can also benefit from label noise in some cases.

In several respects, our theoretical analysis acts a counterpoint to the folklore in this area. For instance, Okamoto & Nobuhiro (1997) analysed the performance of the $k$nn classifier in the presence of label noise. They considered relatively small problem sizes and small values of k, where the $k$nn classifier performs poorly when trained with imperfect labels; on the other hand, our Theorem 2 reveals that for larger values of k, which diverge with n, the asymptotic effect of label noise is relatively modest, and may even improve the performance of the classifier. As another example, Manwani & Sastry (2013) and Ghosh et al. (2015) claim that SVM classifiers perform poorly in the presence of label noise; our Theorem 3 presents a different picture, however, at least as far as the rate of convergence of the excess risk is concerned. Finally, in two-class Gaussian discriminant analysis, Lachenbruch (1966) showed that LDA is robust to homogeneous label noise when the two classes are equally likely (see also Frénay & Verleysen, 2014, Section III-A). We observe, though, that this robustness is very much the exception rather than the rule: if the prior probabilities are not equal, then the LDA classifier is almost invariably not consistent when trained with imperfect labels; cf. Theorem 4.

Although it is not the focus of this paper, we mention briefly that another line of work on label noise investigates techniques for identifying mislabelled observations and either relabelling them, or simply removing them from the training data set. Such methods are sometimes referred to as data cleansing or editing techniques; see for instance Wilson (1972), Wilson & Martinez (2000) and Cheng et al. (2017); as well as Frénay & Kabán (2014, Section 3.2), who provide a general overview of popular methods for editing training data sets. Other authors focus on estimating the noise rates and recovering the clean class-conditional distributions (Blanchard et al., 2016; Northcutt et al., 2017).

The following notation is used throughout the paper. We write $\|\cdot\|$ for the Euclidean norm on $\mathbb{R}^d$, and for $r > 0$ and $z \in \mathbb{R}^d$, write $B_z(r) = \{x \in \mathbb{R}^d : \|x - z\| < r\}$ for the open Euclidean ball of radius r centered at z, and let $a_d = \pi^{d/2}/\Gamma(1 + d/2)$ denote the d-dimensional volume of $B_0(1)$. If $A \in \mathbb{R}^{d\times d}$, we write $\|A\|_{\mathrm{op}}$ for its operator norm. For a sufficiently smooth real-valued function f defined on $D \subseteq \mathbb{R}^m$, and for $x \in D$, we write $\dot{f}(x) = (f_1(x), \ldots, f_m(x))^T$

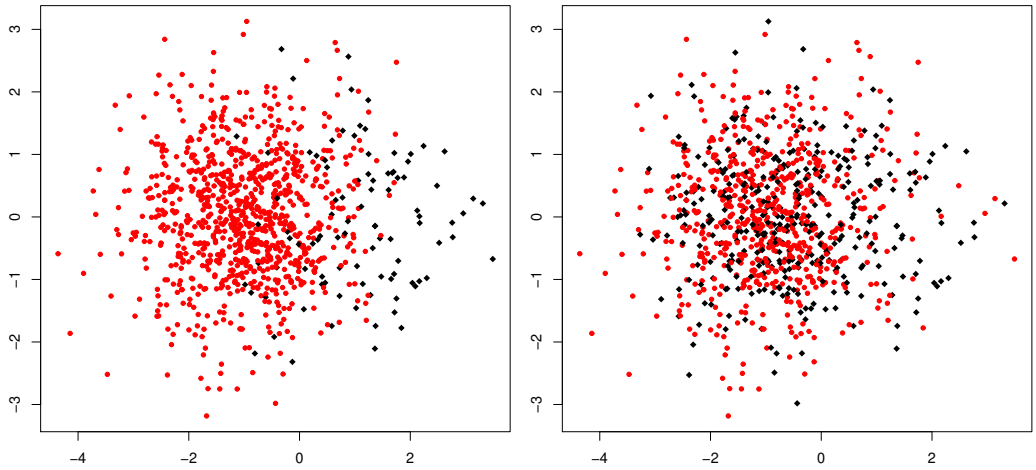

Fig. 1. One training dataset from the model in Example 1 for $n = 1000$, without label noise (left) and with label noise (right). We plot class 0 in red and class 1 in black.

and $\ddot{f}(x) = (f_{jk}(x))_{j,k=1}^{m}$ for its gradient vector and Hessian matrix at x respectively. Finally, we write $\triangle$ for symmetric difference, so that $\mathcal{A} \triangle \mathcal{B} = (\mathcal{A}^c \cap \mathcal{B}) \cup (\mathcal{A} \cap \mathcal{B}^c)$.

We conclude this section with a preliminary study to demonstrate our new results for the $k$nn, SVM and LDA classifiers in the homogeneous noise case.

Example 1. In this motivating example, we demonstrate the surprising effects of imperfect labels on the performance of the $k$nn, SVM and LDA classifiers. We generate n independent training data pairs, where the prior probabilities of classes 0 and 1 are $9/10$ and $1/10$ respectively; class 0 and 1 observations have bivariate normal distributions with means $\mu_0 = (-1, 0)^T$ and $\mu_1 = (1, 0)^T$ respectively, and common identity covariance matrix. We then introduce label noise in the training data set by flipping the true training data labels independently with probability $\rho = 0{\cdot}3$. One example of a data set of size $n = 1000$ from this model, both before and after label noise is added, is shown in Fig. 1.

In Fig. 2, we present the percentage error rates, both with and without label noise, of the $k$nn, SVM and LDA classifiers. The error rates were estimated by the average over 1000 repetitions of the experiment of the percentage of misclassified observations on a test set, without label noise, of size 1000. We set $k = k_n = \lfloor n^{2/3}/2 \rfloor$ for the $k$nn classifier, and set the tuning parameter $\lambda = 1$ for the SVM classifier; see (8).

In this simple setting where the decision boundary of the Bayes classifier is a hyperplane, all three classifiers perform very well with perfectly labelled training data, especially LDA, whose derivation was motivated by Gaussian class-conditional distributions with common covariance matrix. With mislabelled training data, the performance of all three classifiers is somewhat affected, but the $k$nn and SVM classifiers are relatively robust to the label noise, particularly for large n. Indeed, we will show that these classifiers remain consistent in this setting. The gap between the performance of the LDA classifier and that of the Bayes classifier, however, persists even for large n; this again is in line with our theory developed in Theorem 4, where we derive the asymptotic risk of the LDA classifier trained with homogeneous label errors. The limiting risk is given explicitly in terms of the noise rate ρ, the prior probabilities, and the Mahalanobis distance between the two class-conditional distributions.

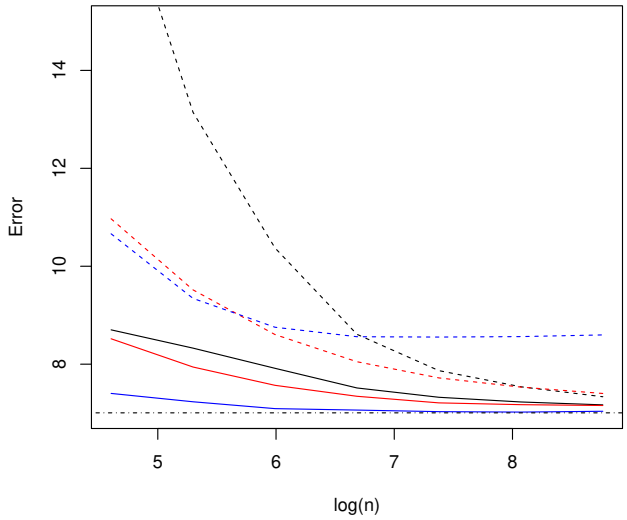

Fig. 2. Risks (%) of the $k$nn (black), SVM (red) and LDA (blue) classifiers trained using perfect (solid lines) and imperfect labels (dotted lines). The dot-dashed line shows the Bayes risk, which is 7·0%.

2. STATISTICAL SETTING

Let $\mathcal{X}$ be a measurable space. In the basic binary classification problem, we observe independent and identically distributed training data pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ taking values in $\mathcal{X} \times \{0, 1\}$ with joint distribution P. The task is to predict the class Y of a new observation X, where $(X, Y) \sim P$ is independent of the training data.

Define the prior probabilities $\pi_1 = \mathrm{pr}(Y = 1) = 1 - \pi_0 \in (0, 1)$ and class-conditional distributions $X \mid \{Y = r\} \sim P_r$ for $r = 0, 1$. The marginal feature distribution of X is denoted P_X and we define the regression function $\eta(x) = \mathrm{pr}\big(Y = 1 \mid X = x\big)$. A classifier C is a measurable function from $\mathcal{X}$ to $\{0, 1\}$, with the interpretation that a point $x \in \mathcal{X}$ is assigned to class $C(x)$.

The risk of a classifier C is $R(C) = \mathrm{pr}\{C(X) \neq Y\}$; it is minimized by the Bayes classifier

$$C^{\mathrm{Bayes}}(x) = \begin{cases} 1 & \text{if } \eta(x) \geq 1/2 \\ 0 & \text{otherwise.} \end{cases}$$

However, since η is typically unknown, in practice we construct a classifier C_n, say, that depends on the n training data pairs. We say (C_n) is consistent if $R(C_n) - R(C^{\mathrm{Bayes}}) \to 0$ as $n \to \infty$. When we write $R(C_n)$ here, we implicitly assume that C_n is a measurable function from $(\mathcal{X} \times \{0, 1\})^n \times \mathcal{X}$ to $\{0, 1\}$, and the probability is taken over the joint distribution of $(X_1, Y_1), \ldots, (X_n, Y_n), (X, Y)$. It is convenient to set $\mathcal{S} = \{x \in \mathcal{X} : \eta(x) = 1/2\}$.

In this paper, we study settings where the true class labels $Y_1, \ldots, Y_n$ for the training data are not observed. Instead we see $\tilde{Y}_1, \ldots, \tilde{Y}_n$, where the noisy label $\tilde{Y}_i$ still takes values in $\{0, 1\}$, but may not be the same as Y_i. The task, however, is still to predict the true class label Y associated with the test point X. We can therefore consider an augmented model where $(X, Y, \tilde{Y}), (X_1, Y_1, \tilde{Y}_1), \ldots, (X_n, Y_n, \tilde{Y}_n)$ are independent and identically distributed triples taking values in $\mathcal{X} \times \{0, 1\} \times \{0, 1\}$.

At this point the dependence between Y and $\tilde{Y}$ is left unrestricted, but we introduce the following notation: define measurable functions $\rho_0, \rho_1 : \mathcal{X} \to [0, 1]$ by $\rho_r(x) = \mathrm{pr}(\tilde{Y} \neq Y \mid X = x, Y = r)$. Thus, letting $Z \mid \{X = x, Y = r\} \sim \mathrm{Bin}(1, 1 - \rho_r(x))$ for $r = 0, 1$, we can write

$\tilde{Y} = ZY + (1-Z)(1-Y)$. We refer to the case where $\rho_0(x) = \rho_1(x) = \rho$ for all $x \in \mathcal{X}$ as ρ-homogeneous noise. Further, let $\tilde{P}$ denote the joint distribution of $(X, \tilde{Y})$, and let $\tilde{\eta}(x) = \mathrm{pr}(\tilde{Y} = 1 \mid X = x)$ denote the regression function for $\tilde{Y}$, so that

$$\begin{aligned}\tilde{\eta}(x) &= \eta(x)\mathrm{pr}(\tilde{Y} = 1 \mid X = x, Y = 1) + \{1 - \eta(x)\}\mathrm{pr}(\tilde{Y} = 1 \mid X = x, Y = 0) \\ &= \eta(x)\{1 - \rho_1(x)\} + \{1 - \eta(x)\}\rho_0(x). \end{aligned} \tag{1}$$

We also define the corrupted Bayes classifier

$$\tilde{C}^{\mathrm{Bayes}}(x) = \begin{cases} 1 & \text{if } \tilde{\eta}(x) \geq 1/2 \\ 0 & \text{otherwise,} \end{cases}$$

which minimizes the corrupted risk $\tilde{R}(C) = \mathrm{pr}\{C(X) \neq \tilde{Y}\}$.

3. EXCESS RISK BOUNDS FOR ARBITRARY CLASSIFIERS

A key property in this work will be that the Bayes classifier is preserved under label noise; more specifically, in Theorem 1(i) below, we will provide conditions under which

$$P_X\big(\{x \in \mathcal{S}^c : \tilde{C}^{\mathrm{Bayes}}(x) \neq C^{\mathrm{Bayes}}(x)\}\big) = 0. \tag{2}$$

In Theorem 1(ii), we go on to show that, under slightly stronger conditions on the label error probabilities and for an arbitrary classifier C, we can bound the excess risk $R(C) - R(C^{\mathrm{Bayes}})$ of predicting the true label by a multiple of the excess risk of predicting a noisy label $\tilde{R}(C) - \tilde{R}(\tilde{C}^{\mathrm{Bayes}})$, where this multiple does not depend on the classifier C. This latter result is particularly useful when the classifier C is trained using the imperfect labels, that is with the training data $(X_1, \tilde{Y}_1), \ldots, (X_n, \tilde{Y}_n)$, because, as will be shown in the next section, we are able to provide further control of $\tilde{R}(C) - \tilde{R}(\tilde{C}^{\mathrm{Bayes}})$ for specific choices of C.

It is convenient to let $\mathcal{B} = \{x \in \mathcal{S}^c : \rho_0(x) + \rho_1(x) < 1\}$, and let

$$\mathcal{A} = \left\{ x \in \mathcal{B} : \frac{\rho_1(x) - \rho_0(x)}{\{2\eta(x) - 1\}\{1 - \rho_0(x) - \rho_1(x)\}} < 1 \right\}.$$

THEOREM 1. *(i) We have*

$$P_X\big(\mathcal{A} \triangle \{x \in \mathcal{B} : \tilde{C}^{\mathrm{Bayes}}(x) = C^{\mathrm{Bayes}}(x)\}\big) = 0. \tag{3}$$

In particular, if $P_X(\mathcal{A}^c \cap \mathcal{S}^c) = 0$, then (2) holds.

(ii) Now suppose, in fact, that there exist $\rho^* < 1/2$ and $a^* < 1$ such that $P_X(\{x \in \mathcal{S}^c : \rho_0(x) + \rho_1(x) > 2\rho^*\}) = 0$, and

$$P_X\left(\left\{ x \in \mathcal{B} : \frac{\rho_1(x) - \rho_0(x)}{\{2\eta(x) - 1\}\{1 - \rho_0(x) - \rho_1(x)\}} > a^* \right\} \right) = 0.$$

Then, for any classifier C,

$$R(C) - R(C^{\mathrm{Bayes}}) \leq \frac{\tilde{R}(C) - \tilde{R}(\tilde{C}^{\mathrm{Bayes}})}{(1 - 2\rho^*)(1 - a^*)}.$$

In Theorem 1(i), the condition $P_X(\mathcal{A}^c \cap \mathcal{S}^c) = 0$ restricts the difference between the two mislabelling probabilities at P_X-almost all $x \in \mathcal{S}^c$, with stronger restrictions where $\eta(x)$ is close to $1/2$ and where $\rho_0(x) + \rho_1(x)$ is close to 1. Moreover, since $\mathcal{A} \subseteq \mathcal{B}$, we also have $P_X(\mathcal{B}^c \cap \mathcal{S}^c) = 0$, which limits the total amount of label noise at each point; cf. Menon et al.

(2016, Assumption 1). In particular, it ensures that

$$\mathrm{pr}(\tilde{Y} \neq Y \mid X = x) = \eta(x)\rho_1(x) + \{1 - \eta(x)\}\rho_0(x) < 1,$$

for P_X-almost all $x \in \mathcal{S}^c$. In part (ii), the requirement on a^* imposes a slightly stronger restriction on the same weighted difference between the two mislabelling probabilities compared with part (i).

The conditions in Theorem 1 generalize those given in the existing literature by allowing a wider class of noise mechanisms. For instance, in the case of ρ-homogeneous noise, we have $P_X(\mathcal{A}^c \cap \mathcal{S}^c) = 0$ provided only that $\rho < 1/2$. In fact, in this setting, we may take $a^* = 0$ (Ghosh et al., 2015, Theorem 1). More generally, we may also take $a^* = 0$ if the noise depends only on the feature vector and not the true class label, i.e. $\rho_0(x) = \rho_1(x)$ for all x (Menon et al., 2016, Proposition 4).

The proof of Theorem 1(ii) relies on the following proposition, which provides a bound on the excess risk for predicting a true label, assuming only that (2) holds.

PROPOSITION 1. *Assume that* (2) *holds. Further, for* $\kappa > 0$*, let*

$$A_\kappa = \Big\{x \in \mathcal{X} : |2\eta(x) - 1| \leq \kappa |2\tilde{\eta}(x) - 1|\Big\}.$$

Then, for any classifier C*,*

$$R(C) - R(C^{\mathrm{Bayes}}) \leq \min\Big[\mathrm{pr}\{C(X) \neq \tilde{C}^{\mathrm{Bayes}}(X)\}, \inf_{\kappa > 0}\Big\{\kappa\{\tilde{R}(C) - \tilde{R}(\tilde{C}^{\mathrm{Bayes}})\} + P_X(A_\kappa^c)\Big\}\Big]. \tag{4}$$

Our main focus in this work is on settings where $\tilde{C}^{\mathrm{Bayes}}$ and C^{Bayes} agree, i.e. (2) holds, because this is where we can hope for classifiers to be robust to label noise. However, in this instance, we present a more general version of Proposition 1 as Proposition A1 in the online supplement; this bounds the excess risk of an arbitrary classifier without the assumption that (2) holds. We see in that result, there is an additional contribution to the risk bound of $R(\tilde{C}^{\mathrm{Bayes}}) - R(C^{\mathrm{Bayes}}) \geq 0$. See also, for instance, Natarajan et al. (2013), who study asymmetric homogeneous noise, where $\rho_0(x) = \rho_0 \neq \rho_1 = \rho_1(x)$, with ρ_0 and ρ_1 known.

We can regard $|2\eta(x) - 1|$ as a measure of the ease of classifying x. Hence, in Proposition 1, we can interpret A_κ as the set of points x where the relative difficulty of classifying x in the corrupted problem compared with its uncorrupted version is controlled. The level of this control can then be traded off against the measure of the exceptional set A_κ^c.

To provide further understanding of Proposition 1, observe that in general, we have

$$\begin{aligned}\tilde{R}(C) - \tilde{R}(\tilde{C}^{\mathrm{Bayes}}) &= \int_{\mathcal{X}} \big[\mathrm{pr}\{C(x) = 0\} - \mathbb{1}_{\{\tilde{\eta}(x) < 1/2\}}\big]\{2\tilde{\eta}(x) - 1\}\, dP_X(x) \\ &\leq \mathrm{pr}\{C(X) \neq \tilde{C}^{\mathrm{Bayes}}(X)\}.\end{aligned}$$

Thus, if $P_X(A_1^c) = 0$, then the second term in the minimum in (4) gives a better bound than the first. However, typically in practice, we would have that $P_X(A_1^c) \neq 0$, and indeed, in Example A1 in the supplementary material, we show that for the 1-nearest neighbour classifier with homogeneous noise, either of the two terms in the minimum in (4) can be smaller, depending on the noise level. As a consequence of Proposition 1, we have the following corollary.

COROLLARY 1. *Suppose that* $(\tilde{C}_n)$ *is a sequence of classifiers satisfying* $\tilde{R}(\tilde{C}_n) \to \tilde{R}(\tilde{C}^{\mathrm{Bayes}})$ *and assume that* (2) *holds. Further, let* $\tilde{\mathcal{S}} = \{x \in \mathcal{X} : \tilde{\eta}(x) = 1/2\}$*. Then*

$$\limsup_{n \to \infty} R(\tilde{C}_n) - R(C^{\mathrm{Bayes}}) \leq P_X(\tilde{\mathcal{S}} \setminus \mathcal{S}).$$

In particular, if $P_X(\tilde{\mathcal{S}} \setminus \mathcal{S}) = 0$*, then* $R(\tilde{C}_n) \to R(C^{\mathrm{Bayes}})$ *as* $n \to \infty$.

The condition $\tilde{R}(\tilde{C}_n) \to \tilde{R}(\tilde{C}^{\mathrm{Bayes}})$ asks that the classifier is consistent for predicting a corrupted test label. In Section 4 we will see that appropriate versions of the corrupted $k$nn and SVM classifiers satisfy this condition, provided, in the latter case, that the feature vectors have compact support. To understand the strength of Corollary 1, consider the special case of ρ-homogeneous noise, and a classifier $\tilde{C}_n$ that is consistent for predicting a noisy label when trained with corrupted data. Then $\tilde{\mathcal{S}} = \mathcal{S}$ by (1), so provided only that $\rho < 1/2$, Corollary 1 ensures that $\tilde{C}_n$ remains consistent for predicting a true label when trained using the corrupted data.

4. ASYMPTOTIC PROPERTIES

4·1. *The k-nearest neighbour classifier*

We now specialize to the case $\mathcal{X} = \mathbb{R}^d$. The $k$nn classifier assigns the test point X to a class based on a majority vote over the class labels of the k nearest points among the training data. More precisely, given $x \in \mathbb{R}^d$, let $(X_{(1)}, Y_{(1)}), \ldots, (X_{(n)}, Y_{(n)})$ be the reordering of the training data pairs such that

$$\|X_{(1)} - x\| \leq \ldots \leq \|X_{(n)} - x\|,$$

where ties are broken by preserving the original ordering of the indices. For $k \in \{1, \ldots, n\}$, the k-nearest neighbour classifier is

$$C^{k\mathrm{nn}}(x) = C_n^{k\mathrm{nn}}(x) = \begin{cases} 1 & \text{if } \frac{1}{k}\sum_{i=1}^{k} \mathbb{1}_{\{Y_{(i)}=1\}} \geq 1/2 \\ 0 & \text{otherwise.} \end{cases}$$

This simple and intuitive method has received considerable attention since it was introduced by Fix & Hodges (1951, 1989). Stone (1977) showed that the $k$nn classifier is universally consistent, i.e., $R(C^{k\mathrm{nn}}) \to R(C^{\mathrm{Bayes}})$ for any distribution P, as long as $k = k_n \to \infty$ and $k/n \to 0$ as $n \to \infty$. For a substantial overview of the early work on the theoretical properties of the $k$nn classifier, see Devroye et al. (1996). Further recent studies include Kulkarni & Posner (1995), Audibert & Tsybakov (2007), Hall et al. (2008), Biau et al. (2010), Samworth (2012), Chaudhuri & Dasgupta (2014), Gadat et al. (2016), Celisse & Mary-Huard (2018) and Cannings et al. (2018).

Here we study the properties of the corrupted k-nearest neighbour classifier

$$\tilde{C}^{k\mathrm{nn}}(x) = \tilde{C}_n^{k\mathrm{nn}}(x) = \begin{cases} 1 & \text{if } \frac{1}{k}\sum_{i=1}^{k} \mathbb{1}_{\{\tilde{Y}_{(i)}=1\}} \geq 1/2 \\ 0 & \text{otherwise,} \end{cases}$$

where $\tilde{Y}_{(i)}$ denotes the corrupted label of $(X_{(i)}, Y_{(i)})$. Since the $k$nn classifier is universally consistent, we have $\tilde{R}(\tilde{C}^{k\mathrm{nn}}) \to \tilde{R}(\tilde{C}^{\mathrm{Bayes}})$ for any choice of k satisfying Stone's conditions. Thus, by Corollary 1, if (2) holds and $P_X(\tilde{\mathcal{S}} \setminus \mathcal{S}) = 0$, then the corrupted $k$nn classifier remains universally consistent. In particular, in the special case of ρ-homogeneous noise, provided only that $\rho < 1/2$, this result tells us that the corrupted $k$nn classifier remains universally consistent.

We now show that, under further regularity conditions on the data distribution P and the noise mechanism, it is possible to give a more precise description of the asymptotic error properties of the corrupted $k$nn classifier. Since our conditions on P, which are slight simplifications of those used in Cannings et al. (2018) to analyse the uncorrupted $k$nn classifier, are a little technical, we give an informal summary of them here, deferring formal statements of our assumptions (A1)–(A4) to just before the proof of Theorem 2 in Section A·2. First, we assume that each of

the class-conditional distributions has a density with respect to Lebesgue measure such that the marginal feature density $\bar{f}$ is continuous and positive. It turns out that the dominant terms in the asymptotic expansion of the excess risk of $k$nn classifiers are driven by the behaviour of P in a neighbourhood $\mathcal{S}^\epsilon$ of the set $\mathcal{S}$, which consists of points that are difficult to classify correctly, so we ask for further regularity conditions on the restriction of P to $\mathcal{S}^\epsilon$. In particular, we ask for both $\bar{f}$ and η to have two well-behaved derivatives in $\mathcal{S}^\epsilon$, and for $\dot{\eta}$ to be bounded away from 0 on $\mathcal{S}$. This amounts to asking that the class-conditional densities, when weighted by the prior probabilities of each class, cut at an angle, and ensures that the set $\mathcal{S}$ is a $(d-1)$-dimensional orientable manifold. Away from the set $\mathcal{S}^\epsilon$, we only require weaker conditions on P_X, and for η to be bounded away from $1/2$. Finally, we ask for two αth moment conditions to hold, namely that $\int_{\mathbb{R}^d} \|x\|^\alpha \, dP_X(x) < \infty$ and $\int_{\mathcal{S}} \bar{f}(x_0)^{d/(\alpha+d)} \, d\mathrm{Vol}^{d-1}(x_0) < \infty$, where $d\mathrm{Vol}^{d-1}$ denotes the $(d-1)$-dimensional volume form on $\mathcal{S}$.

For $\beta \in (0, 1/2)$, let $K_\beta = \{\lceil (n-1)^\beta \rceil, \ldots, \lfloor (n-1)^{1-\beta} \rfloor\}$ denote the set of values of k to be considered for the $k$nn classifier. Define

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

where

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

We will also make use of a condition on the noise rates near the Bayes decision boundary:

Assumption B1. There exist $\delta > 0$ and a function $g : (1/2 - \delta, 1/2 + \delta) \to [0, 1)$ that is differentiable at $1/2$, with the property that for x such that $\eta(x) \in (1/2 - \delta, 1/2 + \delta)$, we have $\rho_0(x) = g(\eta(x))$ and $\rho_1(x) = g(1 - \eta(x))$.

This assumption asks that, when $\eta(x)$ is close to $1/2$, the probability of label noise depends only on x through $\eta(x)$, and moreover, this probability varies smoothly with $\eta(x)$. In other words, Assumption B1 says that the probability of mislabelling an observation with true class label 0 depends only on the extent to which it appeared to be from class 1; conversely, the probability of mislabelling an observation with true label 1 depends only, and in a symmetric way, on the extent to which it appeared to be from class 0. To give just one of many possible examples, one could imagine that the probability that a doctor misdiagnoses a malignant tumour as benign depends on the extent to which it appears to be malignant, and vice versa. We remark that Menon et al. (2016, Definition 11) introduce a related probabilistically transformed noise model, where $\rho_0 = g_0 \circ \eta$ and $\rho_1 = g_1 \circ \eta$, but they also require that g_0 and g_1 are increasing on $[0, 1/2]$ and decreasing on $[1/2, 1]$; see also Bylander (1997).

THEOREM 2. *Assume* A1, A2, A3 *and* A4(α). *Suppose that ρ_0, ρ_1 are continuous, and that both*

$$\rho^* = \frac{1}{2} \sup_{x \in \mathbb{R}^d} \{\rho_0(x) + \rho_1(x)\} < \frac{1}{2}$$

and

$$a^* = \sup_{x \in \mathcal{B}} \frac{\rho_1(x) - \rho_0(x)}{\{2\eta(x) - 1\}\{1 - \rho_0(x) - \rho_1(x)\}} < 1.$$

Moreover, assume B1 *holds with the additional requirement that g is twice continuously differentiable, $\dot{g}(1/2) > 2g(1/2) - 1$ and that $\ddot{g}$ is uniformly continuous. Then we have two cases:*

(i) Suppose that $d \geq 5$ and $\alpha > 4d/(d-4)$. Then for each $\beta \in (0, 1/2)$,

$$R(\tilde{C}^{k\mathrm{nn}}) - R(C^{\mathrm{Bayes}}) = \frac{B_1}{k\{1 - 2g(1/2) + \dot{g}(1/2)\}^2} + B_2\Big(\frac{k}{n}\Big)^{4/d} + o\bigg(\frac{1}{k} + \Big(\frac{k}{n}\Big)^{4/d}\bigg)$$

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

(ii) Suppose that either $d \leq 4$, or, $d \geq 5$ and $\alpha \leq 4d/(d-4)$. Then for each $\beta \in (0, 1/2)$ and each $\epsilon > 0$ we have

$$R(\tilde{C}^{k\mathrm{nn}}) - R(C^{\mathrm{Bayes}}) = \frac{B_1}{k\{1 - 2g(1/2) + \dot{g}(1/2)\}^2} + o\bigg(\frac{1}{k} + \Big(\frac{k}{n}\Big)^{\frac{\alpha}{\alpha+d}-\epsilon}\bigg)$$

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

The proof of Theorem 2 is given in Section A·2, and involves two key ideas. First, we demonstrate that the conditions assumed for η also hold for the corrupted regression function $\tilde{\eta}$. Second, we show that the dominant asymptotic contribution to the desired excess risk $R(\tilde{C}^{k\mathrm{nn}}) - R(C^{\mathrm{Bayes}})$ is $\{\tilde{R}(\tilde{C}^{k\mathrm{nn}}) - \tilde{R}(\tilde{C}^{\mathrm{Bayes}})\}/\{1 - 2g(1/2) + \dot{g}(1/2)\}$, a scalar multiple of the excess risk when predicting a noisy label. We then conclude the argument by appealing to Cannings et al. (2018, Theorem 1), and of course, can recover the conclusion of that result for noiseless labels as a special case of Theorem 2 by setting $g = 0$.

In the conclusion of Theorem 2(i), the terms $B_1/[k\{1 - 2g(1/2) + \dot{g}(1/2)\}^2]$ and $B_2(k/n)^{4/d}$ can be thought of as the leading order contributions to the variance and squared bias of the corrupted $k$nn classifier respectively. It is both surprising and interesting to note that the type of label noise considered here affects only the leading order variance term compared with the noiseless case; the dominant bias term is unchanged. To give a concrete example, ρ-homogeneous noise satisfies the conditions of Theorem 2, and in the setting of Theorem 2(i), we see that the dominant variance term is inflated by a factor of $(1 - 2\rho)^{-2}$.

We now quantify the relative asymptotic performance of the corrupted $k$nn and uncorrupted $k$nn classifiers. Since this performance depends on the choice of k in each case, we couple these choices together in the following way: given any k to be used by the uncorrupted classifier $C^{k\mathrm{nn}}$, and given the function g from Theorem 2, we consider the choice

$$k_g = \big\lfloor \{1 - 2g(1/2) + \dot{g}(1/2)\}^{-2d/(d+4)} k \big\rfloor \tag{5}$$

for the noisy label classifier $\tilde{C}^{k\mathrm{nn}}$. This coupling reflects the ratio of the optimal choices of k for the corrupted and uncorrupted label settings.

COROLLARY 2. *Under the assumptions of Theorem 2(i), and provided that $B_2 > 0$, we have that for any $\beta \in (0, 1/2)$,*

$$\frac{R(\tilde{C}^{k_g\mathrm{nn}}) - R(C^{\mathrm{Bayes}})}{R(C^{k\mathrm{nn}}) - R(C^{\mathrm{Bayes}})} \to \{1 - 2g(1/2) + \dot{g}(1/2)\}^{-8/(d+4)}, \tag{6}$$

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

If $\dot{g}(1/2) > 2g(1/2)$, then the limiting regret ratio in (6) is less than 1 – this means that the label noise helps in terms of the asymptotic performance! This is due to the fact that, under the noise model in Theorem 2, if $\dot{g}(1/2) > 2g(1/2)$ then for points X_i with $\eta(X_i)$ close to $1/2$, the noisy labels $\tilde{Y}_i$ are more likely than the true labels Y_i to be equal to the Bayes labels, $\mathbb{1}_{\{\eta(X_i) \geq 1/2\}}$. To understand this phenomenon, first note that by rearranging (1), we have

$$\tilde{\eta}(x) - 1/2 = \{\eta(x) - 1/2\}\{1 - \rho_0(x) - \rho_1(x)\} + \frac{1}{2}\{\rho_0(x) - \rho_1(x)\}.$$

Thus $\tilde{\eta}(x) - 1/2 = \eta(x) - 1/2$ for $x \in \mathcal{S}$ using B1. On the other hand, for $x \in \mathcal{S}^c$, we have

$$\tilde{\eta}(x) - 1/2 = \{\eta(x) - 1/2\}\left(1 - \rho_0(x) - \rho_1(x) + \frac{\rho_0(x) - \rho_1(x)}{2\eta(x) - 1}\right). \tag{7}$$

We next study the term in the second parentheses on the right-hand side above. Write $t = \eta(x) - 1/2$. Then, for x such that $|\eta(x) - 1/2| \in (0, \delta)$, we have $\rho_0(x) = g(1/2 + t)$ and $\rho_1(x) = g(1/2 - t)$. It follows, for such x, that

$$\begin{aligned} 1 - \rho_0(x) - \rho_1(x) + \frac{\rho_0(x) - \rho_1(x)}{2\eta(x) - 1} &= 1 - g(1/2 + t) - g(1/2 - t) + \frac{g(1/2 + t) - g(1/2 - t)}{2t} \\ &\to 1 - 2g(1/2) + \dot{g}(1/2) \end{aligned}$$

as $|t| \searrow 0$. Since $1 - 2g(1/2) + \dot{g}(1/2) > 1$, we obtain that for any $\varepsilon \in \big(0, \dot{g}(1/2)/2 - g(1/2)\big)$, there exists $\delta_0 \in (0, \delta)$ such that for all x with $|\eta(x) - 1/2| \in (0, \delta_0)$, we have that

$$1 - \rho_0(x) - \rho_1(x) + \frac{\rho_0(x) - \rho_1(x)}{2(\eta(x) - 1/2)} > 1 - 2g(1/2) + \dot{g}(1/2) - \varepsilon > 1.$$

This together with (7) ensures that, for all x such that $|\eta(x) - 1/2| \in (0, \delta_0)$, we have

$$|\tilde{\eta}(x) - 1/2| > |\eta(x) - 1/2|.$$

Example 2. Suppose that for some $g_0 \in (0, 1/2)$ and $h_0 > 2 - 1/g_0$ we have $g(1/2 + t) = g_0(1 + h_0 t)$ for $t \in (-\delta, \delta)$. Then $g(1/2) = g_0$ and $\dot{g}(1/2) = g_0 h_0$, which gives $1 - 2g(1/2) + g'(1/2) = 1 + (h_0 - 2)g_0$. We therefore see from Corollary 2 that if $h_0 < 2$, then the limiting regret ratio is greater than 1, but if $h_0 > 2$, then the limiting regret ratio is less than one, so the label noise aids performance.

4·2. *Support vector machine classifiers*

In general, the term support vector machines (SVM) refers to classifiers of the form

$$C^{\mathrm{SVM}}(x) = C_n^{\mathrm{SVM}}(x) = \begin{cases} 1 & \text{if } \hat{f}(x) \geq 0 \\ 0 & \text{otherwise,} \end{cases}$$

where the decision function $\hat{f}$ satisfies

$$\hat{f} \in \operatorname*{argmin}_{f \in H}\left\{\frac{1}{n}\sum_{i=1}^{n} L(Y_i, f(X_i)) + \Omega(\lambda, \|f\|_H)\right\}.$$

See, for example, Cortes & Vapnik (1995) and Steinwart & Christmann (2008). Here $L : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is a loss function, $\Omega : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is a regularization function, $\lambda > 0$ is a tuning parameter and H is a reproducing kernel Hilbert space with norm $\|\cdot\|_H$ (Steinwart & Christmann, 2008, Chapter 4).

We focus throughout on the L1-SVM, where $L(y, t) = \max\{0, 1 - (2y - 1)t\}$ is the hinge loss function and $\Omega(\lambda, t) = \lambda t^2$. Let $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be the positive definite kernel function associated with the reproducing kernel Hilbert space. We consider the Gaussian radial basis function, namely $K(x, x') = \exp(-\sigma^2\|x - x'\|^2)$, for $\sigma > 0$. The corrupted SVM classifier is

$$\tilde{C}^{\mathrm{SVM}}(x) = \tilde{C}_n^{\mathrm{SVM}}(x) = \begin{cases} 1 & \text{if } \tilde{f}(x) \geq 0 \\ 0 & \text{otherwise,} \end{cases} \tag{8}$$

where

$$\tilde{f} \in \operatorname*{argmin}_{f \in H} \biggl\{ \frac{1}{n} \sum_{i=1}^{n} \max\{0, 1 - (2\tilde{Y}_i - 1) f(X_i)\} + \lambda \|f\|_H^2 \biggr\}. \tag{9}$$

Steinwart (2005, Corollary 3.6 and Example 3.8) show that the uncorrupted L1-SVM classifier is consistent as long as P_X is compactly supported and $\lambda = \lambda_n$ is such that $\lambda_n \to 0$ but $n\lambda_n / (|\log \lambda_n|^{d+1}) \to \infty$. Therefore, under these conditions, provided that (2) holds and $P_X(\tilde{\mathcal{S}} \setminus \mathcal{S}) = 0$, by Corollary 1, we have that $R(\tilde{C}^{\mathrm{SVM}}) \to R(C^{\mathrm{Bayes}})$ as $n \to \infty$.

Under further conditions on the noise probabilities and the distribution P, we can also provide more precise control of the excess risk for the SVM classifier. Our analysis will make use of the results in Steinwart & Scovel (2007), who study the rate of convergence of the SVM classifier with Gaussian kernels in the noiseless label setting. Other works of note on the rate of convergence of SVM classifiers include Lin (1999) and Blanchard et al. (2008); see also Steinwart & Christmann (2008, Chapters 6 and 8).

We recall two definitions used in the perfect labels context. The first of these is the well-known margin assumption of, for example, Audibert & Tsybakov (2007). We say that the distribution P satisfies the margin assumption with parameter $\gamma_1 \in [0, \infty)$ if there exists $\kappa_1 > 0$ such that

$$P_X(\{x \in \mathbb{R}^d : 0 < |\eta(x) - 1/2| \leq t\}) \leq \kappa_1 t^{\gamma_1}$$

for all $t > 0$. If P satisfies the margin assumption for all $\gamma_1 \in [0, \infty)$ then we say P satisfies the margin assumption with parameter ∞. The margin assumption controls the probability mass of the region where η is close to $1/2$.

The second definition we need is that of the geometric noise exponent (Steinwart & Scovel, 2007, Definition 2.3). Let $\mathcal{S}_+ = \{x \in \mathbb{R}^d : \eta(x) > 1/2\}$ and $\mathcal{S}_- = \{x \in \mathbb{R}^d : \eta(x) < 1/2\}$, and for $x \in \mathbb{R}^d$, let $\tau_x = \inf_{x' \in \mathcal{S} \cup \mathcal{S}_+} \|x - x'\| + \inf_{x' \in \mathcal{S} \cup \mathcal{S}_-} \|x - x'\|$. We say that the distribution P has geometric noise exponent $\gamma_2 \in [0, \infty)$ if there exists $\kappa_2 > 0$, such that

$$\int_{\mathbb{R}^d} |2\eta(x) - 1| \exp\Bigl(-\frac{\tau_x^2}{t^2}\Bigr) \, dP_X(x) \leq \kappa_2 t^{\gamma_2 d}$$

for all $t > 0$. If P has geometric noise exponent γ_2 for all $\gamma_2 \in [0, \infty)$ then we say it has geometric noise exponent ∞.

Under these two conditions, Steinwart & Scovel (2007, Theorem 2.8) show that, if P_X is supported on the closed unit ball, then for appropriate choices of the tuning parameters, the SVM classifier achieves a convergence rate of $O(n^{-\Gamma+\epsilon})$ for every $\epsilon > 0$, where

$$\Gamma = \begin{cases} \frac{\gamma_2}{2\gamma_2+1} & \text{if } \gamma_2 \leq \frac{\gamma_1+2}{2\gamma_1} \\ \frac{2\gamma_2(\gamma_1+1)}{2\gamma_2(\gamma_1+2)+3\gamma_1+4} & \text{otherwise.} \end{cases}$$

In the imperfect labels setting, and under our stronger assumption on the noise mechanism when η is close to $1/2$, we see that the SVM classifier trained with imperfect labels satisfies the same bound on the rate of convergence as in the perfect labels case.

THEOREM 3. *Suppose that P satisfies the margin assumption with parameter $\gamma_1 \in [0, \infty]$, has geometric noise exponent $\gamma_2 \in (0, \infty)$ and that P_X is supported on the closed unit ball. Assume the conditions of Theorem* 1*(ii) and B1 holds. Then*

$$R(\tilde{C}^{\mathrm{SVM}}) - R(C^{\mathrm{Bayes}}) = O(n^{-\Gamma+\epsilon}),$$

as $n \to \infty$, for every $\epsilon > 0$. If $\gamma_2 = \infty$, then the same conclusion holds provided $\sigma_n = \sigma$ is a constant with $\sigma > 2d^{1/2}$.

4·3. *Linear discriminant analysis*

If $P_0 = N_d(\mu_0, \Sigma)$ and $P_1 = N_d(\mu_1, \Sigma)$, then the Bayes classifier is

$$C^{\mathrm{Bayes}}(x) = \begin{cases} 1 & \text{if } \log\left(\frac{\pi_1}{\pi_0}\right) + \left(x - \frac{\mu_1+\mu_0}{2}\right)^T \Sigma^{-1}(\mu_1 - \mu_0) \geq 0 \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

The Bayes risk can be expressed in terms of π_0, π_1, and the squared Mahalanobis distance $\Delta^2 = (\mu_1 - \mu_0)^T \Sigma^{-1} (\mu_1 - \mu_0)$ between the classes as

$$R(C^{\mathrm{Bayes}}) = \pi_0 \Phi\left(\frac{1}{\Delta}\log\left(\frac{\pi_1}{\pi_0}\right) - \frac{\Delta}{2}\right) + \pi_1 \Phi\left(\frac{1}{\Delta}\log\left(\frac{\pi_0}{\pi_1}\right) - \frac{\Delta}{2}\right),$$

where Φ denotes the standard normal distribution function.

The LDA classifier is constructed by substituting training data estimates of $\pi_0, \pi_1, \mu_0, \mu_1,$ and Σ in to (10). With imperfect training data labels, and for $r = 0, 1$, we define estimates $\hat{\pi}_r = n^{-1} \sum_{i=1}^n \mathbb{1}_{\{\tilde{Y}_i = r\}}$ of π_r, as well as estimates $\hat{\mu}_r = \sum_{i=1}^n X_i \mathbb{1}_{\{\tilde{Y}_i = r\}} / \sum_{i=1}^n \mathbb{1}_{\{\tilde{Y}_i = r\}}$ of the class-conditional means μ_r, and set

$$\hat{\Sigma} = \frac{1}{n-2} \sum_{i=1}^n \sum_{r=0}^1 (X_i - \hat{\mu}_r)(X_i - \hat{\mu}_r)^T \mathbb{1}_{\{\tilde{Y}_i = r\}}.$$

This allows us to define the corrupted LDA classifier

$$\tilde{C}^{\mathrm{LDA}}(x) = \tilde{C}_n^{\mathrm{LDA}}(x) = \begin{cases} 1 & \text{if } \log\left(\frac{\hat{\pi}_1}{\hat{\pi}_0}\right) + \left(x - \frac{\hat{\mu}_1+\hat{\mu}_0}{2}\right)^T \hat{\Sigma}^{-1}(\hat{\mu}_1 - \hat{\mu}_0) \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

Consider now the ρ-homogeneous noise setting. In this case, writing $\tilde{P}_r$, $r \in \{0, 1\}$, for the distribution of $X \mid \{\tilde{Y} = r\}$, we have $\tilde{P}_r = p_r N_d(\mu_r, \Sigma) + (1 - p_r) N_d(\mu_{1-r}, \Sigma)$, where $p_r = \pi_r(1-\rho)/\{\pi_r(1-\rho) + \pi_{1-r}\rho\}$. Notice that while $\hat{\pi}_r, \hat{\mu}_r$ and $\hat{\Sigma}$ are intended to be estimators of π_r, μ_r and Σ, respectively, with label noise these will in fact be consistent estimators of $\tilde{\pi}_r = \pi_r(1-\rho) + \pi_{1-r}\rho$, $\tilde{\mu}_r = p_r \mu_r + (1 - p_r)\mu_{1-r}$, and $\tilde{\Sigma} = \Sigma + \alpha(\mu_1 - \mu_0)(\mu_1 - \mu_0)^T$, respectively, where $\alpha > 0$ is given in the proof of Theorem 4.

We will also make use of the following well-known lemma in the homogeneous label noise case (e.g. Ghosh et al., 2015, Theorem 1), which holds for an arbitrary classifier and data generating distribution. We include the short proof for completeness.

LEMMA 1. *For ρ-homogeneous noise with $\rho \in [0, 1/2)$ and for any classifier C, we have $R(C) = \{\tilde{R}(C) - \rho\}/(1 - 2\rho)$. Moreover, $R(C) - R(C^{\mathrm{Bayes}}) = \{\tilde{R}(C) - \tilde{R}(C^{\mathrm{Bayes}})\}/(1 - 2\rho)$.*

The following is the main result of this subsection.

THEOREM 4. *Suppose that $P_r = N_d(\mu_r, \Sigma)$ for $r = 0, 1$ and that the noise is ρ-homogeneous with $\rho \in [0, 1/2)$. Then*

$$\lim_{n\to\infty} \tilde{C}^{\mathrm{LDA}}(x) = \begin{cases} 1 & \text{if } c_0 + \left(x - \frac{\mu_1+\mu_0}{2}\right)^T \Sigma^{-1}(\mu_1 - \mu_0) > 0 \\ 0 & \text{if } c_0 + \left(x - \frac{\mu_1+\mu_0}{2}\right)^T \Sigma^{-1}(\mu_1 - \mu_0) < 0, \end{cases}$$

where

$$c_0 = \left\{(1-2\rho) + \frac{\rho(1-\rho)(1+\pi_0\pi_1\Delta^2)}{(1-2\rho)\pi_1\pi_0}\right\} \log\left(\frac{(1-2\rho)\pi_1 + \rho}{(1-2\rho)\pi_0 + \rho}\right) - \frac{(\pi_1 - \pi_0)\rho(1-\rho)\Delta^2}{2\{(1-2\rho)^2\pi_1\pi_0 + \rho(1-\rho)\}}.$$

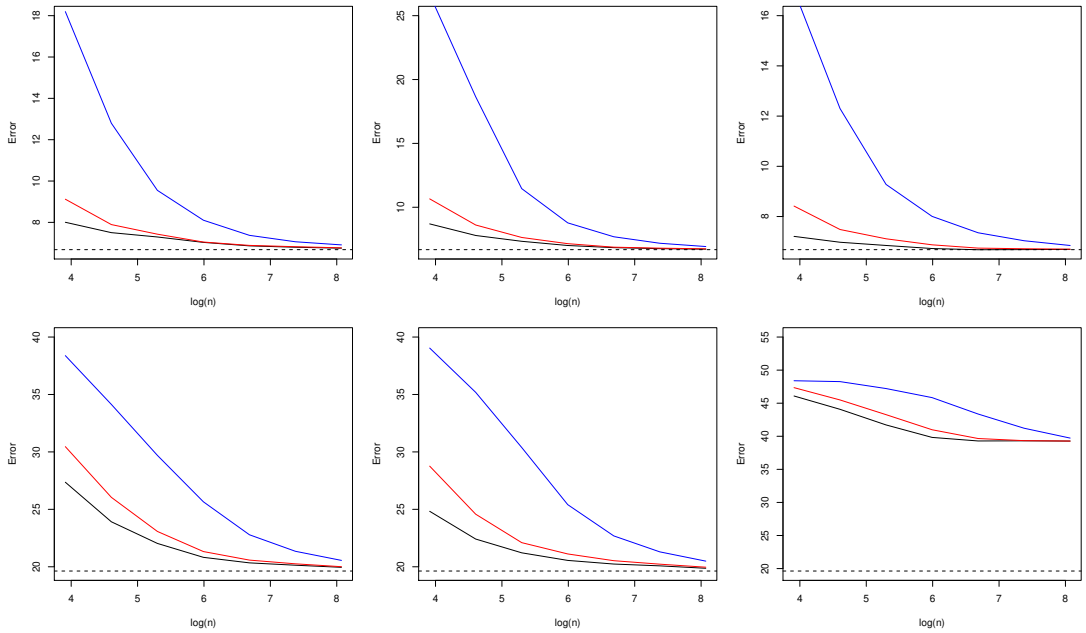

Fig. 3. Risk estimates for the $k$nn (left), SVM (middle) and LDA (right) classifiers. Top: Model 1, $d = 2$, $\pi_1 = 0{\cdot}5$, Bayes risk = 6·68%, shown as the black dotted line. Bottom: Model 2, $d = 2$, Bayes risk = 19·63%. We present the results without label noise (black) and with homogeneous label noise at rate $\rho = 0{\cdot}1$ (red) and $0{\cdot}3$ (blue).

As a consequence,

$$\lim_{n\to\infty} R(\tilde{C}^{\mathrm{LDA}}) = \pi_0 \Phi\left(\frac{c_0}{\Delta} - \frac{\Delta}{2}\right) + \pi_1 \Phi\left(-\frac{c_0}{\Delta} - \frac{\Delta}{2}\right) \geq R(C^{\mathrm{Bayes}}). \tag{11}$$

For each $\rho \in (0, 1/2)$ and $\pi_0 \neq \pi_1$, there exists a unique value of $\Delta > 0$ for which equality in the inequality in (11) *is attained.*

The first conclusion of this theorem reveals the interesting fact that, regardless of the level $\rho \in (0, 1/2)$ of label noise, the limiting corrupted LDA classifier has a decision hyperplane that is parallel to that of the Bayes classifier; see also Lachenbruch (1966) and Manwani & Sastry (2013, Corollary 1). However, for each fixed $\rho \in (0, 1/2)$ and $\pi_0 \neq \pi_1$, there is only one value of $\Delta > 0$ for which the offset is correct and the corrupted LDA classifier is consistent.

5. NUMERICAL COMPARISON

In this section, we investigate empirically how the different types of label noise affect the performance of the k-nearest neighbour, support vector machine and linear discriminant analysis classifiers. We consider two different model settings for the pair (X, Y):

Model 1: Let $\mathrm{pr}(Y = 1) = \pi_1 \in \{0{\cdot}5,\ 0{\cdot}9\}$ and $X \mid \{Y = r\} \sim N_d(\mu_r, I_d)$, where $\mu_1 = (3/2, 0, \ldots, 0)^T = -\mu_0 \in \mathbb{R}^d$ and I_d denotes the d by d identity matrix.

Model 2: For $d \geq 2$, let $X \sim U([0,1]^d)$ and $\mathrm{pr}(Y = 1 \mid X = x) = \eta(x_1, \ldots, x_d) = \min\{4(x_1 - 1/2)^2 + 4(x_2 - 1/2)^2, 1\}$.

In each setting, our risk estimates are based on an uncorrupted test set of size 1000, and we repeat each experiment 1000 times. This ensures that all standard errors are less than 0·4% and 0·14 for the risk and regret ratio estimates, respectively; in fact, they are often much smaller.

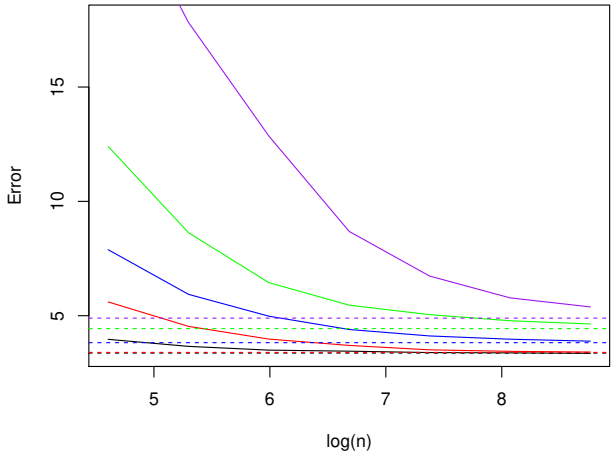

Fig. 4. Risk estimates for the LDA classifier. Model 1, $d = 5$, $\pi_1 = 0{\cdot}9$, Bayes risk = 3·37%. We present the estimated error without label noise (black) and with homogeneous label noise at rate $\rho = 0{\cdot}1$ (red), 0·2 (blue), 0·3 (green) and 0·4 (purple). The dotted lines show the corresponding asymptotic limit as given by Theorem 4.

Our first goal is to illustrate numerically our consistency and inconsistency results for the $k$nn, SVM and LDA classifiers. In Fig. 3 we present estimates of the risk for the three classifiers with different levels of homogeneous label noise. We see that for Model 1 when the class prior probabilities are equal, all three classifiers perform well and in particular appear to be consistent, even when as many as 30% of the training data labels are incorrect on average. For the $k$nn and SVM classifiers we observe very similar results for Model 2; the LDA classifier does not perform well in this setting, however, since the Bayes decision boundary is non-linear. These conclusions are in accordance with Corollary 1 and Theorem 4.

We further investigate the effect of homogeneous label noise on the performance of the LDA classifier for data from Model 1, but now when $d = 5$ and the class prior probabilities are unbalanced. Recall that in Theorem 4 we derived the asymptotic limit of the risk in terms of the Mahalanobis distance between the true class distributions, the class prior probabilities and the noise rate. In Fig. 4, we present the estimated risks of the LDA classifier for data from Model 1 with $\pi_1 = 0{\cdot}9$ for different homogeneous noise rates alongside the limit as specified by Theorem 4. This articulates the inconsistency of the corrupted LDA classifier, as observed in Theorem 4.

Finally, we study empirically the asymptotic regret ratios for the $k$nn and SVM classifiers. We focus on the noise model in Example 2 in Section 4, where the label errors occur at random as follows: fix $g_0 \in (0, 1/2)$, $h_0 > 2 - 1/g_0$, we let $g(1/2 + t) = \max[0, \min\{g_0(1 + h_0 t), 2g_0\}]$, then set $\rho_0(x) = g(\eta(x))$ and $\rho_1(x) = g(1 - \eta(x))$. In particular, we use the following settings: (i) $g_0 = 0{\cdot}1$, $h_0 = 0$; (ii) $g_0 = 0{\cdot}1$, $h_0 = -1$; (iii) $g_0 = 0{\cdot}1$, $h_0 = 1$; (iv) $g_0 = 0{\cdot}1$, $h_0 = 2$; (v) $g_0 = 0{\cdot}1$, $h_0 = 3$. Noise setting (i), where $h_0 = 0$, corresponds to g_0-homogeneous noise.

For the $k$nn classifier, where k is chosen to satisfy the conditions of Corollary 2, our theory says that when $d = 5$ in Models 1 and 2, the asymptotic regret ratios in the five noise settings are 1·22, 1·37, 1·10, 1 and 0·92 respectively. We see from the left-hand plots of Fig. 5 that, for k chosen separately in the corrupted and uncorrupted cases via cross-validation, the empirical results provide good agreement with our theory, especially in the last three settings. Reasons for the slight discrepancies between our asymptotic theory and empirically observed regret ratios in the first two noise settings include the following facts: the choices of k in the noisy and noiseless

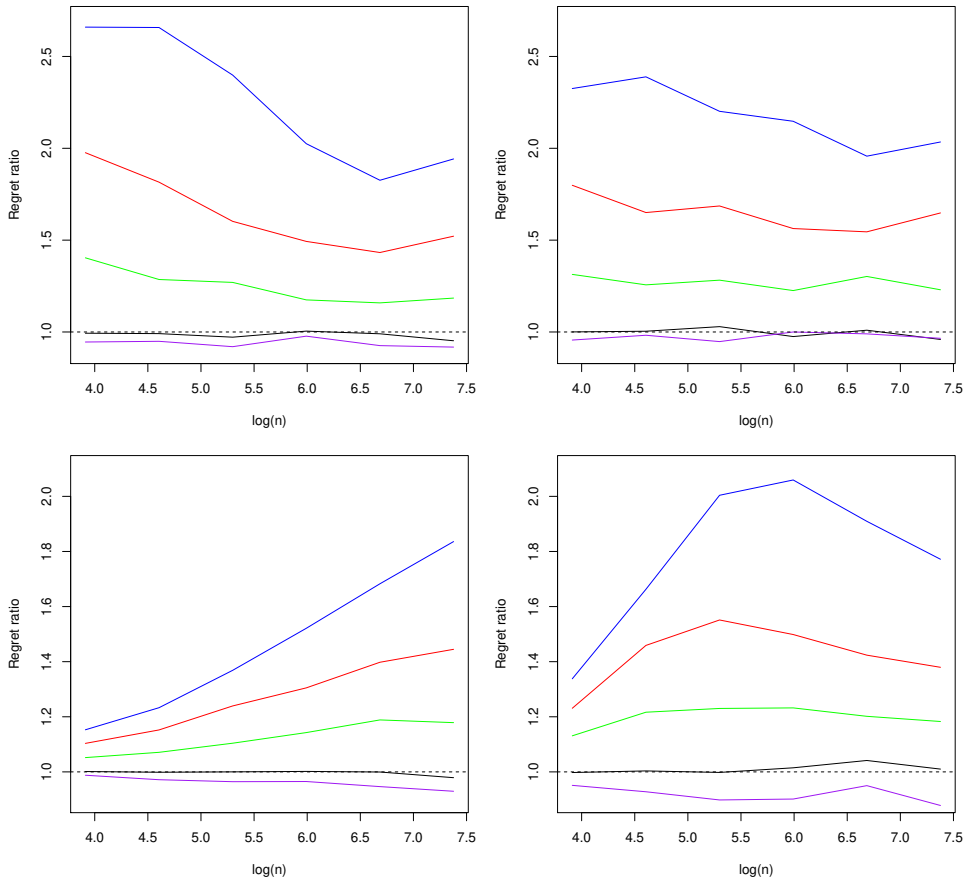

Fig. 5. Estimated regret ratios for the $k$nn (left) and SVM (right) classifiers. Top: Model 1, with $d = 5$ and $\pi_1 = 0{\cdot}5$. Bottom: Model 2, with $d = 5$. We present the results with label noise of type (i – red), (ii – blue), (iii – green), (iv – black), and (v – purple).

label settings do not necessarily satisfy (5) exactly; the asymptotics in n may not have fully 'kicked in'; and Monte Carlo error – when n is large, we are computing the ratio of two small quantities, so the standard error tends to be larger. The performance of the SVM classifier is similar to that of the $k$nn classifier for both models.

Finally, we discuss tuning parameter selection. We have seen that for the $k$nn classifier the the choice of k is important for achieving the optimal bias–variance trade-off; see also Hall et al. (2008). Similarly, we need to choose an appropriate value of λ for the SVM classifier; in practice, this is typically done via cross-validation. When the classifier $\tilde{C}$ is trained with ρ-homogeneous noisy labels, we would like to select a tuning parameter to minimize $R(\tilde{C})$, but since the training data is corrupted, a tuning parameter selection method will target the minimizer of $\tilde{R}(\tilde{C})$. However, by Lemma 1, we have that $R(\tilde{C}) = \{\tilde{R}(\tilde{C}) - \rho\}/(1 - 2\rho)$, and it follows that our tuning parameter selection method requires no modification when trained with noisy labels. In the heterogeneous noise case, however, we do not have this direct relationship; see Inouye et al. (2017) for more on this topic.

In our simulations, we chose k for the $k$nn classifier and λ for the SVM classifier via leave-one-out and 10-fold cross-validation respectively, where the cross-validation was performed over

the noisy training dataset. Moreover, for the SVM classifier, we used the default choice $\sigma^2 = 1/d$ for the hyper-parameter for the kernel function.

ACKNOWLEDGEMENTS

The authors would like to thank Jinchi Lv for introducing us to this topic, and the anonymous reviewers for helpful and constructive comments. The second author is partly supported by an National Science Foundation CAREER Award, and the third author is supported by an Engineering and Physical Sciences Research Council Fellowship and a grant from the Leverhulme Trust. The authors would like to thank the Isaac Newton Institute for Mathematical Sciences for support and hospitality during the programme 'Statistical Scalability' when work on this paper was undertaken.

SUPPLEMENTARY MATERIAL

Supplementary material available at *Biometrika* online includes the proofs of our theoretical results and an illustrative example.

REFERENCES

Angluin, D. & Laird, P. (1988) Learning from noisy examples. *Mach. Learn.*, **2**, 343–370.

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

Awasthi, P., Balcan, M.-F., Urner, R. & Haghtalab, N. (2015) Efficient learning of linear separators under bounded noise. *Proc. Mach. Learn. Res.*, **4**, 167–190.

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

Blanchard, G, Bousquet, O. & Massart, P. (2008) Statistical performance of support vector machines. *Ann. Statist.*, **36**, 489–531.

Blanchard, G., Flaska, M., Handy, G., Pozzi, S. & Scott, C. (2016) Classification with asymmetric label noise: consistency and maximal denoising. *Electron. J. Statist*, **10**, 2780–2824.

Bolton, R. J. & Hand, D. J. (2002) Statistical fraud detection: a review. *Statistical Science (with discussion)*, **17**, 235–255.

Bootkrajang, J. (2016) A generalised label noise model for classification in the presence of annotation errors. *Neurocomputing*, **192**, 61–71.

Bootkrajang, J. & Kabán, A. (2012) Label-noise robust logistic regression and its applications. In: *Machine Learning and Knowledge Discovery in Databases*, (Eds: Flach, P. A., De Bie, T. & Cristianini, N.), Springer Berlin Heidelberg, Berlin, **1**, 143–158.

Bootkrajang, J. & Kabán, A. (2014) Learning kernel logistic regression in the presence of label noise. *Pattern Rocognition*, **47**, 3641–3655.

Bylander, T. (1997) Learning probabilistically consistent linear threshold functions. *COLT*, 62-71.

Cannings, T. I., Berrett, T. B. & Samworth, R. J. (2018) Local nearest neighbour classification with applications to semi-supervised learning. *ArXiv e-prints*, 1704.00642.

Celisse, A. & Mary-Huard, T. (2018) Theoretical analysis of cross-validation for estimating the risk of the k-nearest neighbor classifier. *J. Mach. Learn. Res.*, **19** 1–54.

Chaudhuri, K. & Dasgupta, S. (2014) Rates of convergence for nearest neighbor classification. *Advances in Neural Information Processing Systems*, **27**, 3437–3445.

Chen, X., Lin, Q. & Zhou, D. (2015) Statistical decision making for optimal budget allocation in crowd labeling. *J. Mach. Learn. Res.* **16**, 1–46.

Cheng, J., Liu, T., Ramamohanarao, K. & Tao, D. (2017) Learning with bounded instance- and label-dependent label noise. *ArXiv e-prints*, 1709.03768.

Cortes, C. & Vapnik, V. (1995) Support vector networks. *Machine Learning*, **20**, 273–297.

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

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

Fix, E. & Hodges, J. L. (1989) Discriminatory analysis – nonparametric discrimination: Consistency properties. *Internat. Statist. Rev.*, **57**, 238–247.

Frénay, B. & Kabán, A. (2014) A comprehensive introduction to label noise. *Proc. Euro. Sym. Artificial Neural Networks*, 667–676.

Frénay, B. & Verleysen, M. (2014) Classification in the presence of label noise: a survey. *IEEE Trans. on NN and Learn. Sys.*, **25**, 845–869.

Gadat, S., Klein, T. & Marteau, C. (2016) Classification with the nearest neighbour rule in general finite dimensional spaces. *Ann. Statist.*, **44**, 982–1001.

Ghosh, A., Manwani, N. & Sastry, P. S. (2015) Making risk minimization tolerant to label noise. *Neurocomputing*, **160**, 93–107.

Hall, P., Park, B. U. & Samworth, R. J. (2008) Choice of neighbour order in nearest-neighbour classification. *Ann. Statist.*, **36**, 2135–2152.

Inouye, D. I., Ravikumar, P., Das, P. & Dutta, A. (2017) Hyperparameter selection under localized label noise via corrupt validation. *NIPS 2017*.

Kulkarni, S. R. & Posner, S. E. (1995) Rates of convergence of nearest neighbor estimation under arbitrary sampling. *IEEE Trans. Info. Th.*, **41**, 1028–1039.

Lara, Ó. D. & Labrador, M. A. (2013) A survey on human activity recognition using wearable sensors. *IEEE Comm. Sur. and Tutor.*, **15**, 1192–1209.

Lachenbruch, P. A. (1966) Discriminant analysis when the initial samples are misclassified. *Technometrics*, **8**, 657–662.

Lachenbruch, P. A. (1974) Discriminant analysis when the initial samples are misclassified ii: Non-random misclassification models. *Technometrics*, **16**, 419–424.

Li, Y., Yang, J., Song, Y., Cao, L., Luo, J. & Li, L.-J. (2017) Learning from noisy labels with distillation. *IEEE Intern. Conf. Comp. Vis.*, 1910–1918.

Lin, Y. (1999) Support vector machines and the Bayes rule in classification. *Dept. of Statist., U. Wisconsin*, Technical report No. 1014. Available at `https://pdfs.semanticscholar.org/8b78/66d7d1e8fb87eb3a061bdedf8c7840947f0d.pdf`

Liu, T. & Tao, D. (2016) Classification with noisy labels by importance reweighting. *IEEE Trans. Pattern Anal. and Mach. Int.*, **38**, 447–461.

Long, P. M. & Servedio, R. A. (2010) Random classification noise defeats all convex potential boosters. *Mach. Learn.*, **78**, 287-304.

Lugosi, G. (1992) Learning with an unreliable teacher. *Pattern Recognition*, **25**, 79–87.

Manwani, N. & Sastry, P S. (2013) Noise tolerance under risk minimization. *IEEE Trans. on Cybernetics*, **43**, 1146–1151.

Menon, A. K., van Rooyen, B. & Natarajan, N. (2016) Learning from binary labels with instance-dependent corruption. *ArXiv e-prints*, 1605.00751.

Natarajan, N., Dhillon, I. S., Ravikumar, P. K. & Tewari, A. (2013) Learning with noisy labels. *NIPS*, **26**, 1196–1204.

Northcutt, C. G., Wu, T. & Chuang, I. L. (2017) Learning with confident examples: Rank pruning for robust classification with noisy labels. *Uncertainty in Artificial Intelligence 2017*, ArXiv:1705.01936.

Okamoto, S. & Nobuhiro, Y. (1997) An average-case analysis of the k-nearest neighbor classifier for noisy domains. in *Proc. 15th Int. Joint Conf. Artif. Intell.*, **1**, 238–243.

Patrini, G., Nielsen, F., Nock, R. & Carioni, M. (2016) Loss factorization, weakly supervised learning and label noise robustness. *ICML 2016*, 708–717.

Patrini, G., Rozza, A., Menon, A. K., Nock, R. & Qu, L. (2017) Making deep neural networks robust to label noise: a loss correction approach. *IEEE Conf. Comp. Vis. & Patt. Recog.*, 1944–1952.

Rolnick, D., Veit, A., Belongie, S. & Shavit, N. (2017) Deep learning is robust to massive label noise. *ArXiv e-prints*, 1705.10694.

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

Schölkopf, B., Herbrich, R. & Smola, A. J. (2001) A generalized representer theorem. In *Proc. 14th Annual Conf. Computational Language Theory*, **2111**, 416–426.

Scott, C., Blanchard, G. & Handy, G. (2013) Classification with asymmetric label noise: consistency and maximal denoising. *JMLR: W&CP*, **30**, 1-23.

Steinwart, I. (2005) Consistency of support vector machines and other regularized kernel classifiers. *IEEE Trans. Inf. Th.*, **51**, 128–142.

Steinwart, I. & Christmann, A. (2008) *Support Vector Machines.* Springer, New York.

Steinwart, I. & Scovel, C. (2007) Fast rates for support vector machines using Gaussian kernels. *Ann. Statist.*, **35**, 575–607.

Stempfel, G. & Ralaivola, L. (2009) Learning SVMs from sloppily labeled data. In *Proc. of the 19th Int. Conf. Artificial Neural Networks*, **1**, 884–893.

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

van Rooyen, B., Menon, A. K. & Williamson, R. C. (2015) Learning with symmetric label noise: the importance of being unhinged. *NIPS*, **28**, 10–18.

Wilson, D. L. (1972) Asymptotic properties of nearest neighbour rules using edited data. *IEEE Trans. on Sys., Man., Cybern.*, **2**, 408–421.

Wilson, D. L. & Martinez, T. R. (2000) Reduction techniques for instance based learning algorithms. *Mach. Learn.*, **38**, 257–286.

Wright, C. F., Fitzgerald, T. W., Jones, W. D., Clayton, S., McRae, J. F., van Kogelenberg, M., King, D. A., Ambridge, K., Barrett, D. M., Bayzetinova, T., Bevan, A. P., Bragin, E., Chatzimichali, E. A., Gribble, S., Jones, P., Krishnappa, N., Mason, L. E., Miller, R., Morley, K. I., Parthiban, V., Prigmore, E., Rajan, D., Sifrim, A., Swaminathan, G. J., Tivey, A. R., Middleton, A., Parker, M., Carter, N. P., Barrett, J. C., Hurles, M. E., Fitzpatrick, D. R. & Firth, H. V. (2015) Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. *The Lancet*, **385**, 1305–1314.

Zhang, Y., Chen, X., Zhou, D. & Jordan, M. I. (2016) Spectral methods meet EM: A provably optimal algorithm for crowdsourcing. *J. Mach. Learn. Res.* **17**, 1–44.