Fast rates for a kNN classifier robust to unknown asymmetric label noise

Reeve, Henry W. J.; Kaban, Ata

Fast Rates for a kNN Classifier Robust to Unknown Asymmetric Label Noise

Henry W. J. Reeve [1] **Ata Kabán** [1]

Abstract

We consider classification in the presence of class-dependent asymmetric label noise with unknown noise probabilities. In this setting, identifiability conditions are known, but additional assumptions were shown to be required for finite sample rates, and so far only the parametric rate has been obtained. Assuming these identifiability conditions, together with a measure-smoothness condition on the regression function and Tsybakov's margin condition, we show that the Robust kNN classifier of Gao et al. attains, the mini-max optimal rates of the *noise-free setting*, up to a log factor, even when trained on data with unknown asymmetric label noise. Hence, our results provide a solid theoretical backing for this empirically successful algorithm. By contrast the standard kNN is not even consistent in the setting of asymmetric label noise. A key idea in our analysis is a simple kNN based method for estimating the maximum of a function that requires far less assumptions than existing mode estimators do, and which may be of independent interest for noise proportion estimation and randomised optimisation problems.

1. Introduction

Label noise is a pervasive issue in real-world classification tasks, as perfectly accurate labels are often very costly, and sometimes impossible, to produce (Natarajan et al., 2018; Cannings et al., 2018; Blanchard et al., 2016; Frénay & Verleysen, 2014).

We consider asymmetric label noise with unknown class-conditional noise probabilities – that is, the labels we observe have randomly flipped in some proportion that depends on the class. This type of noise is both realistic and amenable to analysis (Blanchard et al., 2016; Natarajan et al., 2018). In this setting the classical kNN algorithm is no longer consistent (see Section 5). Most existing theoretical work in this direction assumes that the noise probabilities are known in advance by the learner (Natarajan et al., 2013), at least approximately (Natarajan et al., 2018). However, in many situations such knowledge is not available, for instance in positive unlabelled (PU) learning (Elkan & Noto, 2008) one may regard unlabelled data as a class of negative examples contaminated with positives in an unknown proportion. Other examples include the problem of nucleur particle classification discussed by (Blanchard et al., 2016). That work also established identifiability conditions sufficient for recovering unknown noise probabilities from corrupted data.

Blanchard et al. (2010) proved that the identifiability conditions are insufficient to obtain finite sample convergence rates. Consequently, Scott et al. (2013) introduced additional conditions external to the classification task with which it is possible to obtain the parametric rate (of order $n^{-1/2}$ where n is the sample size) (Blanchard et al., 2016). To the best of our knowledge it is unknown if faster rates are possible with unknown asymmetric label noise.

Here we answer this question in the affirmative by analysing an existing Robust kNN classifier (Gao et al., 2018). Previously, Gao et al. (2018) conducted a comprehensive empirical study which demonstrates that the Robust kNN, introduced therein, typically outperforms a range of competitors for classification problems with asymmetric label noise. Gao et al. (2018) also proved the consistency of their method, but only under the restrictive assumption of prior knowledge of the label noise probabilities.

We prove that the Robust kNN classifier attains fast rates for classification problems in a flexible non-parametric setting with unknown asymmetric label noise. More precisely, we work under a measure smoothness condition on the regression function, introduced in recent analyses of kNN in the noise-free setting (Chaudhuri & Dasgupta, 2014), termed the 'modified Lipschitz' condition in (Döring et al., 2017), as well as Tsybakov's margin condition. We assume in addition conditions equivalent to those of label noise identifiability (Blanchard et al., 2016; Menon et al., 2015). We show that the Robust kNN introduced by (Gao et al., 2018) attains, up to a log factor, the known minimax optimal fast

[1]School of Computer Science, University of Birmingham, UK. Correspondence to: Henry W. J. Reeve <henry-wjreeve@gmail.com>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

rate of the label noise free setting – despite the presence of unknown asymmetric label noise.

2. Problem Setup

Suppose we have a feature space $\mathcal{X}$ with a metric ρ and a set of labels $\mathcal{Y} = \{0, 1\}$. Let $\mathbb{P}$ be a fixed but unknown distribution on $\mathcal{X} \times \mathcal{Y}$. Our goal is to learn a classifier $\phi : \mathcal{X} \to \mathcal{Y}$ which minimises the risk

$$\mathcal{R}(\phi) := \mathbb{E}[\phi(X) \neq Y].$$

Our data will be generated by a *corrupted* distribution $\mathbb{P}_{\text{corr}}$ on $\mathcal{X} \times \mathcal{Y}$ with asymmetric label noise, so there exist probabilities $p_0, p_1 \in (0, 1)$ with $p_0 + p_1 < 1$ such that random pairs $(X, \tilde{Y}) \sim \mathbb{P}_{\text{corr}}$ are generated by $(X, Y) \sim \mathbb{P}$ and $\tilde{Y} \neq Y$ with probability p_Y and $\tilde{Y} = Y$ otherwise, i.e. $p_0 = \mathbb{P}_{\text{corr}}[\tilde{Y} = 1|Y = 0]$ and $p_1 = \mathbb{P}_{\text{corr}}[\tilde{Y} = 0|Y = 1]$. We have access to a data set $\mathcal{D}_{\text{corr}} = \{(X_i, \tilde{Y}_i)\}_{i\in[n]}$ only, consisting of i.i.d. pairs generated from the corrupted distribution $(X_i, \tilde{Y}_i) \sim \mathbb{P}_{\text{corr}}$.

We let μ denote the marginal distribution over the features i.e. $\mu(A) = \mathbb{P}[X \in A]$ for Borel sets $A \subseteq \mathcal{X}$, and let $\eta : \mathcal{X} \to [0, 1]$ denote the regression function i.e. $\eta(x) = \mathbb{P}[Y = 1|X = x]$. Further, let $\mathcal{X}_\mu \subseteq \mathcal{X}$ denote the support of the measure μ. It follows from the assumption of feature independent label noise that the corrupted distribution $\mathbb{P}_{\text{corr}}$ has the same marginal distribution as $\mathbb{P}$ i.e. $\mathbb{P}_{\text{corr}}[X \in A] = \mu(A)$ for $A \subseteq \mathcal{X}$. Denote by $\eta_{\text{corr}} : \mathcal{X} \to [0, 1]$ the corrupted regression function $\eta_{\text{corr}}(x) = \mathbb{P}_{\text{corr}}[\tilde{Y} = 1|X = x]$. As observed in (Menon et al., 2015), η_{corr} and η are related by

$$\begin{aligned}\eta_{\text{corr}}(x) &= (1 - p_1) \cdot \mathbb{P}[Y = 1|X = x] + p_0 \cdot \mathbb{P}[Y = 0|X = x] \\ &= (1 - p_0 - p_1) \cdot \eta(x) + p_0. \end{aligned} \quad (1)$$

We shall use this connection to provide a label noise robust plug-in classifier.

3. Approach – roadmap

The 'plug-in' classification method is inspired by the fact that the mapping $\phi_* : \mathcal{X} \to \mathcal{Y}$ for $x \in \mathcal{X}$ defined by $\phi_*(x) = \mathbb{1}\{\eta(x) \geq 1/2\}$ is a Bayes classifier and minimises the risk $\mathcal{R}(\phi)$ over all measurable classifiers $\phi : \mathcal{X} \to \mathcal{Y}$. The approach is to first produce an estimate $\hat{\eta} : \mathcal{X} \to [0, 1]$ of the regression function η and then take $\hat{\phi}(x) := \mathbb{1}\{\hat{\eta}(x) \geq 1/2\}$. To apply this method in the label-noise setting we must first give a method for constructing an estimate $\hat{\eta}$ based upon the corrupted sample $\mathcal{D}_{\text{corr}}$. By eq. (1) for each $x \in \mathcal{X}$ we have

$$\eta(x) = (1 - p_0 - p_1)^{-1} \cdot (\eta_{\text{corr}}(x) - p_0). \quad (2)$$

However, all quantities on the LHS are unknown. Our strategy is to decompose the problem under mild conditions, so that we can plug in estimates. The following simple lemma makes this precise.

Lemma 3.1. *Let $\hat{\eta}_{corr} : \mathcal{X} \to [0, 1]$ be an estimate of η_{corr} and define $\hat{\eta} : \mathcal{X} \to [0, 1]$ by $\hat{\eta}(x) := (\hat{\eta}_{corr}(x) - \hat{p}_0) / (1 - \hat{p}_0 - \hat{p}_1)$. Suppose that $p_0 + p_1 < 1$, and $\hat{p}_0, \hat{p}_1 \in [0, 1)$ with $\hat{p}_0 + \hat{p}_1 < 1$. Suppose further that $\max\{|\hat{p}_0 - p_0|, |\hat{p}_1 - p_1|\} \leq (1 - p_0 - p_1)/4$. Then for all $x \in \mathcal{X}$ we have*

$$\begin{aligned}&|\hat{\eta}(x) - \eta(x)| \leq \dots \\ &8 \cdot \frac{\max\{|\hat{\eta}_{corr}(x) - \eta_{corr}(x)|, |\hat{p}_0 - p_0|, |\hat{p}_1 - p_1|\}}{1 - p_0 - p_1}.\end{aligned}$$

Proof. The lemma follows from eq. (1) by a straightforward manipulation. See Appendix B for details. □

We note that the preconditions in Lemma 3.1 that involve estimates may be ensured by a sufficient sample sizes, and are therefore not restrictive.

Consequently, we shall obtain $\hat{\eta}(x)$ in two steps, summarised in the plug-in template Algorithm 1. First we construct an estimator $\hat{\eta}_{\text{corr}}$ for the corrupted regression function η_{corr} based upon the corrupted sample $\mathcal{D}_{\text{corr}}$ using supervised regression methods. The key remaining challenge is then to obtain estimates $\hat{p}_0$ and $\hat{p}_1$ for p_0 and p_1, respectively. The latter is known to be impossible without further assumptions (see Section 4 in (Scott et al., 2013)). Next, we discuss the

Algorithm 1 Plug-in classification with label noise

1. Compute an estimate $\hat{\eta}_{\text{corr}}$ of the corrupted regression function η_{corr} based on $\mathcal{D}_{\text{corr}}$;
2. Compute $\hat{p}_0$ and $\hat{p}_1$ by estimating the extrema of $\hat{\eta}_{\text{corr}}$;
3. Let $\hat{\phi}(x) := \mathbb{1}\{\hat{\eta}_{\text{corr}}(x) \geq 1/2 \cdot (1 + \hat{p}_0 - \hat{p}_1)\}$.

assumptions that we employ for the remainder of this work.

3.1. Main Assumptions and Relation to Previous Work

We employ two kinds of assumptions: (i) Assumptions 1 and 2 represent identifiability conditions for asymmetric label noise; (ii) Assumptions 3 and 4 are conditions under which minimax optimal fast rates are known in the noise-free setting. We now briefly explain each of these, which also serves to place our forthcoming analysis into the context of previous work.

We already made use of the following in Lemma 3.1:

Assumption 1 (Most labels are correct)**.** $p_0 + p_1 < 1$.

Assumption 2 (Range assumption)**.** *We have* $\inf_{x\in\mathcal{X}_\mu}\{\eta(x)\} = 0$ *and* $\sup_{x\in\mathcal{X}_\mu}\{\eta(x)\} = 1$.

Assumption 2 was introduced by Menon et al. (2015) who showed it to be equivalent to the 'mutual irreducibility' condition given in (Scott et al., 2013; Blanchard et al., 2016).

The above form will be more directly useful, since from Assumption 2 and eq. (1) it follows that $\inf_{x\in\mathcal{X}_\mu}\{\eta_{\text{corr}}(x)\} = p_0$ and $\sup_{x\in\mathcal{X}_\mu}\{\eta_{\text{corr}}(x)\} = 1-p_1$. Hence, we may obtain estimates $\hat{p}_0$ and $\hat{p}_1$ by estimating the extrema of the corrupted regression function η_{corr}.

Recall that the above assumptions alone do not permit finite sample convergence rates; therefore Scott (2015) assumed in addition that there are positive measure balls B_0, B_1 in the input space such that $\forall x \in B_i\ \eta(x) = i$, and obtained the parametric rate, i.e. of order $n^{-1/2}$. We will not assume this, instead we now consider assumptions that have already succeeded in establishing fast rates in the noise-free setting.

The following smoothness condition with respect to the marginal distribution was first proposed by Chaudhuri & Dasgupta (2014), and further employed by Döring et al. (2017) who also termed it the 'modified Lipschitz' condition. It generalises the combination of Hölder-continuity and strong density assumptions (i.e. marginal density bounded from below) that have been prevalent in previous theoretical analyses of plug-in classifiers after (Audibert et al., 2007).

Definition 3.1 (Measure-smoothness)**.** *A function $f : \mathcal{X} \to [0,1]$ is* measure-smooth *with exponent $\lambda > 0$ and constant $\omega > 0$ if for all $x_0, x_1 \in \mathcal{X}$ we have $|f(x_0) - f(x_1)| \leq \omega \cdot \mu\left(B_{\rho(x_0,x_1)}(x_0)\right)^\lambda$.*

Assumption 3 (Measure smooth regression function)**.** *The regression function η is measure-smooth with exponent λ and constant ω.*

A sufficient condition for Assumption 3 to hold with $\lambda = \beta/d$ is that η is β-Hölder and μ is absolutely continuous with respect to the Lebesgue measure on $[0,1]^d$ with a uniform lower bound on the density. However, Assumption 3 does not require the existence of a density for μ and also applies naturally to classification in general metric spaces, including discrete distributions (Döring et al., 2017).

The final assumption is Tsybakov's margin condition, which has become a widely used device for explaining fast rates, since it was first proposed by Mammen & Tsybakov (1999).

Assumption 4 (Tsybakov margin condition)**.** *There exists $\alpha \geq 0$ and $C_\alpha \geq 1$ such that for all $\xi > 0$ we have*

$$\mu\left(\left\{x \in \mathcal{X} : 0 < \left|\eta(x) - \frac{1}{2}\right| < \xi\right\}\right) \leq C_\alpha \cdot \xi^\alpha.$$

Note that Assumption 4 always holds with $\alpha = 0$ and $C_\alpha = 1$; hence it is not restrictive.

Under the assumptions of Tsybakov margin, and measure-smoothness of the regression function, Chaudhuri & Dasgupta (2014) obtained, in the noise-free setting (i.e. without label noise), for the k nearest neighbor (kNN) classifier, the convergence rate of order $n^{-\frac{\lambda(\alpha+1)}{2\lambda+1}}$ – which corresponds (after having made explicit the dimensional dependence) to the minimax optimal rate computed by Audibert et al. (2007) (that is, the lower bound is over all classifiers). With these two distributional assumptions, this rate can therefore be regarded as quantifying the statistical hardness of classification in the minimax sense (see e.g. (Tsybakov, 2008)) when a perfect labelling is available. It is not at all obvious whether, and how, the same rate can be achieved in the presence of unknown asymmetric label noise conditions? The aim in the remainder of this paper is to answer this question.

3.2. Notation and tools

Whilst we are motivated by the estimation of η_{corr} we shall frame our results in a more general fashion for clarity. Suppose we have a distribution $\mathbb{P}$ on $\mathcal{X} \times [0,1]$ and let $f : \mathcal{X} \to [0,1]$ be the function $f(x) := \mathbb{E}[Z|X = x]$. Our goals are to estimate f and its extrema based on a sample $\mathcal{D}_f = \{(X_i, Z_i)\}_{i\in[n]}$ with $(X_i, Z_i) \sim \mathbb{P}$ generated i.i.d. We let $\mathbf{X} := \{X_i\}_{i\in[n]}$ and $\mathbf{Z} := \{Z_i\}_{i\in[n]}$.

Given $x \in \mathcal{X}$ we define $\{\tau_{n,q}(x)\}_{q\in[n]}$ to be an enumeration of $[n]$ such that for each $q \in [n-1]$, $\rho\left(x, X_{\tau_{n,q}(x)}\right) \leq \rho\left(x, X_{\tau_{n,q+1}(x)}\right)$. We define the k-nearest neighbour estimate $\hat{f}_{n,k} : \mathcal{X} \to [0,1]$ of f by

$$\hat{f}_{n,k}(x) := \frac{1}{k} \cdot \sum_{q\in[k]} Z_{\tau_{n,q}(x)}. \tag{3}$$

Given a point $x \in \mathcal{X}$ and $r > 0$ we let $B_r(x)$ denote the open metric ball of radius r, centered at x. It will be useful to give a high probability bound on the measure of an open metric ball centered at a given point with random radius equal to the distance of its k-th nearest neighbour.

Lemma 3.2. *Take $x \in \mathcal{X}, k \in [n]$ and $\zeta \geq 0$. Then,*

$$\mathbb{P}^n\left[\mu\left(B_{\rho\left(x, X_{\tau_{n,k}(x)}\right)}(x)\right) > \frac{(1+\zeta)k}{n}\right] \leq e^{-k(\zeta - \log(1+\zeta))}.$$

A bound of this form appears in (Biau & Devroye, 2015) (Sec. 1.2) for the special case where the marginal distribution has a continuous distribution function. Their proof relies of the fact that, in this special case $\mu(B_{\rho\left(x, X_{\tau_{n,k}(x)}\right)}(x))$ follows the distribution of the k-th uniform order statistic whose properties are well studied. However, since we consider a general metric space setting and do not assume a continuous distribution function, below we show from first principles that this bound is still valid, by exploiting the continuity properties of measures.

Proof. For any $x \in \mathcal{X}$ and $p \in [0,1]$ we define (following Chaudhuri & Dasgupta (2014)) the smallest radius for which

the open ball centered at x has probability at least p:

$$r_p(x) = \inf\{r > 0 : \mu(B_r(x)) \geq p\}.$$

Take $r > r_p(x)$, so $\mu(B_r(x)) \geq p$. Note that $\rho\left(x, X_{\tau_{n,k}(x)}\right) \geq r$ if and only if $\sum_{i\in[n]} \mathbb{1}_{\{X_i \in B_r(x)\}} < k$. Moreover, taking

$$\tilde{p} = \frac{1}{n}\sum_{i\in[n]} \mathbb{E}\left[\mathbb{1}_{\{X_i\in B_r(x)\}}\right] = \mu(B_r(x)) \geq p,$$

implies $k \leq (1-\epsilon)n\tilde{p} \leq n\tilde{p}$ where $\epsilon = 1 - k/(np)$. Thus, by the multiplicative Chernoff bound – Theorem 4.5 in (Mitzenmacher & Upfal, 2005) – we have,

$$\begin{aligned}
\mathbb{P}^n\left[\rho\left(x, X_{\tau_{n,k}(x)}\right) \geq r\right] &= \mathbb{P}^n\left[\sum_{i\in[n]} \mathbb{1}_{\{X_i\in B_r(x)\}} < k\right] \\
&\leq \mathbb{P}^n\left[\sum_{i\in[n]} \mathbb{1}_{\{X_i\in B_r(x)\}} < (1-\epsilon)n\tilde{p}\right] \\
&\leq \exp(-n\tilde{p}[\epsilon + (1-\epsilon)\log(1-\epsilon)]) \\
&\leq \exp(-np[\epsilon + (1-\epsilon)\log(1-\epsilon)]).
\end{aligned}$$

Since the above inequality holds for all $r > r_p(x)$, it follows by continuity of μ from above that we have

$$\mathbb{P}^n\left[\rho\left(x, X_{\tau_{n,k}(x)}\right) > r_p(x)\right] \leq e^{-np[\epsilon+(1-\epsilon)\log(1-\epsilon)]}.$$

This implies that with probability at least $1 - \exp(-np[\epsilon + (1-\epsilon)\log(1-\epsilon)])$ we have

$$\mu\left(B_{\rho\left(x, X_{\tau_{n,k}(x)}\right)}(x)\right) \leq \mu(B_{r_p(x)}) \leq p,$$

where the last inequality follows by continuity of measure from below.

Recall that $\epsilon = 1 - k/(np)$. To obtain the conclusion of the lemma we first note that the bound holds trivially whenever $\zeta \geq n/k - 1$ since this implies

$$\begin{aligned}
&\mathbb{P}^n\left[\mu\left(B_{\rho\left(x, X_{\tau_{n,k}(x)}\right)}(x)\right) > \frac{(1+\zeta)k}{n}\right] \\
&\quad\leq \mathbb{P}^n\left[\mu\left(B_{\rho\left(x, X_{\tau_{n,k}(x)}\right)}(x)\right) > 1\right] = 0.
\end{aligned}$$

For $\zeta \in [0, n/k - 1]$, we choose $p = (1+\zeta)k/n \in [0,1]$. Plugging into ϵ yields $\epsilon = \zeta/(1+\zeta)$, and rearranging the r.h.s. of the probability bound completes the proof. □

When the centre of the metric ball is one of the data points, we have the following.

Corollary 3.1. *Take $k, j \in [n]$ and $\zeta > 0$. Then,*

$$\mathbb{P}^n\left[\mu(B_{\rho\left(X_j, X_{\tau_{n,k}(x)}\right)}(X_j)) > \frac{(1+\zeta)k}{n}\right] \leq e^{-(k-1)(\zeta-\log(1+\zeta))}.$$

Since this is a bound also for the ball with a non-random centre point, we may use it in both cases.

Proof. Fix X_j and apply Lemma 3.2 to the $k-1$-th nearest neighbour in the remaining sample of size $n-1$. Expectation w.r.t. X_j on both sides leaves the RHS unchanged. Then use that $\frac{k-1}{n-1} < \frac{k}{n}$. □

4. Results

We are now in a position to follow through the plan of our analysis. The following three subsections correspond directly to the three steps of the algorithm template (see Algorithm 1) through a kNN based classification rule.

4.1. Pointwise function estimation

As a first step we deal with point-wise estimation of the corrupted regression function, which we approach as a kNN regression task (Kpotufe, 2011; Jiang, 2019). However, for our purposes we require a bound that holds both for non-random points $x \in \mathcal{X}_\mu$ and for feature vectors X_j occurring in the data, for reasons that will become clear in the subsequent Section 4.2 where we will need to estimate the maximum of the function.

Theorem 4.1 (Pointwise estimation bound)**.** *Suppose that f satisfies measure-smoothness with exponent $\lambda > 0$ and constant ω. Take $n \in \mathbb{N}$, $\delta \in (0,1)$, $k \in \mathbb{N} \cap [4\log(3/\delta) + 1, n/2]$ and suppose that $\mathcal{D}_f$ is generated i.i.d. from $\mathbb{P}$. Suppose that X is either a fixed point $x \in \mathcal{X}_\mu$ or X_j for some fixed $j \in [n]$. The following bound holds with probability at least $1-\delta$ over $\mathcal{D}_f$*

$$\left|\hat{f}_{n,k}(X) - f(X)\right| \leq \sqrt{\frac{\log(3/\delta)}{2k}} + \omega\cdot\left(\frac{2k}{n}\right)^{\lambda}.$$

We leave k unspecified for now, but we see that a choice of $k \in \Theta(\omega^{-\frac{2}{2\lambda+1}} \cdot n^{\frac{2\lambda}{2\lambda+1}})$ gives the rate $\mathcal{O}(\omega^{\frac{1}{2\lambda+1}} \cdot n^{-\frac{\lambda}{2\lambda+1}})$.

The following lemma will be handy.

Lemma 4.1. *Let f, X, n, k as in Theorem 4.1, and $q \in [k]$. Then for any $\delta > 0$, and $n/2 \geq k \geq 4\log(1/\delta) + 1$, we have w.p. $1-\delta$ that*

$$\left|f\left(X_{\tau_{n,q}(X)}\right) - f(X)\right| \leq \omega\cdot\left(\frac{2k}{n}\right)^{\lambda}.$$

Proof. By the measure-smoothness property,

$$\begin{aligned}
\left|f\left(X_{\tau_{n,q}(X)}\right) - f(X)\right| &\leq \omega\cdot\mu\left(B_{\rho(X, X_{\tau_{n,q}(x)})}(X)\right)^{\lambda} \\
&\leq \omega\cdot\mu\left(B_{\rho(X, X_{\tau_{n,k}(X)})}(X)\right)^{\lambda} \\
&\leq \omega\cdot\left(\frac{(1+\zeta)k}{n}\right)^{\lambda},
\end{aligned}$$

with probability $1 - \exp(-(k-1)(\zeta - \log(1+\zeta)))$, where $\zeta \geq 0$, and the last inequality follows by Corollary 3.1.

For simplicity we choose $\zeta = 1$, whence $1/(1-\log(2)) < 4$, so for $k \geq 4\log(1/\delta)+1$ we have with probability $1-\delta$ the statement of the Lemma. Finally, note that the measure of the ball cannot be larger than 1, so we require $k \leq n/2$. □

Proof. We shall use the notation $\tilde{f}_{n,k}(x) = \mathbb{E}_{\mathbf{Z}|\mathbf{X}}\left[\hat{f}_{n,k}(x)\right] = \frac{1}{k}\cdot\sum_{q\in[k]} f\left(X_{\tau_{n,q}(x)}\right)$.

By the triangle inequality we have

$$\left|\hat{f}_{n,k}(X) - f(X)\right| \leq \left|\hat{f}_{n,k}(X) - \tilde{f}_{n,k}(X)\right| + \left|\tilde{f}_{n,k}(X) - f(X)\right|. \tag{4}$$

To bound the first term, note that X is a deterministic function of $\mathbf{X} = \{X_i\}_{i\in[n]}$. By construction we have $\hat{f}_{n,k}(X) = \frac{1}{k}\cdot\sum_{q\in[k]} Z_{\tau_{n,q}(X)}$. Moreover, for each $q \in [k]$, $Z_{\tau_{n,q}(X)}$ is a random variable in $[0,1]$ with conditional expectation (given $\mathbf{X}$) equal to $\mathbb{E}_{\mathbf{Z}|\mathbf{X}}\left[Z_{\tau_{n,q}(X)}\right] = f\left(X_{\tau_{n,q}(X)}\right)$, and these are independent. Hence, it follows from Chernoff bounds (Boucheron et al., 2013) that the first term is bounded with probability at least $1-2\delta/3$ over $\mathcal{D}_f$, as the following

$$\left|\hat{f}_{n,k}(X) - \tilde{f}_{n,k}(X)\right| \leq \sqrt{\frac{\log(3/\delta)}{2k}}. \tag{5}$$

To bound the second term in eq. (4) we use Lemma 4.1 with $1-\delta/3$, so for the allowed range of values of k we have

$$\left|\tilde{f}_{n,k}(X) - f(X)\right| \leq \frac{1}{k}\cdot\sum_{q\in[k]}\left|f\left(X_{\tau_{n,q}(X)}\right) - f(X)\right| \tag{6}$$

$$\leq \omega\cdot\left(\frac{2k}{n}\right)^{\lambda}. \tag{7}$$

Taking the union bound, eqs. (5) and (6) hold simultaneously with probability at least $1-\delta$. Plugging inequalities (5) and (6) back into (4) completes the proof. □

4.2. Maximum estimation with kNN

In Section 3 we discussed how the noise probabilities p_0 and p_1 are determined by the extrema of the corrupted regression function η_{corr}. This motivates the question of determining the maximum of a function f, which is the focus of this section, although we believe the results of this section may also be of independent interest. As in Section 4.1, we shall assume we have access to a sample $\mathcal{D}_f = \{(X_i, Z_i)\}_{i\in[n]}$ with $(X_i, Z_i) \sim \mathbb{P}$ generated i.i.d. where $\mathbb{P}$ is an unknown distribution on $\mathcal{X}\times[0,1]$ with $f(x) = \mathbb{E}[Z|X=x]$. Our aim is to estimate $M(f) := \sup_{x\in\mathcal{X}_\mu}\{f(x)\}$ based on $\mathcal{D}_f$. We give a bound for a simple estimator under the assumption of measure-smoothness of the regression function.

Before proceeding we should point out that mode-estimation via kNN was previously proposed by (Dasgupta & Kpotufe, 2014; Jiang & Kpotufe, 2017; Jiang, 2019), but both solve a related but different problem to ours. The former papers deal with the unsupervised problem of finding the point where the density is highest. The latter work deals with finding the point where a function is maximal. The key difference is that performance is judged in terms of distance in the input space, whereas we care about distance in function output. As a consequence, previous works require strong curvature assumptions: That the Hessian exists and is negative definite for all modal points. By contrast, we are able to work on metric spaces where the notion of a Hessian does not even make sense, and we only require the measure-smoothness condition which holds, for instance, whenever the regression function is Hölder and its density is bounded from below. We also do not require a bounded input domain.

Take the following estimator for $M(f)$, defined as the empirical maximum of the values of the regression estimator:

$$\widehat{M}_{n,k}(f) = \max_{i\in[n]}\left\{\widehat{f}_{n,k}(X_i)\right\}.$$

Theorem 4.2 (Maximum estimation bound with measure-smoothness)**.** *Suppose that f is measure-smoothness with exponent $\lambda > 0$ and constant $\omega > 0$. Take $n \in \mathbb{N}$, $\delta \in (0,1)$, $k \in \mathbb{N}\cap[4\log(2/\delta)+1, n/2]$. Suppose further that $\mathcal{D}_f$ is generated i.i.d. from $\mathbb{P}$. Then for any $\delta \in (0,1)$ the following holds with probability at least $1-\delta$ over $\mathcal{D}_f$,*

$$\left|\widehat{M}_{n,k}(f) - M(f)\right| \leq \sqrt{\frac{\log(6n/\delta)}{2k}} + 2\omega\cdot\left(\frac{2k}{n}\right)^{\lambda}.$$

Proof. By Theorem 4.1 combined with the union bound, the following holds simultaneously for all $i \in [n]$, with probability at least $1-\delta/2$ over $\mathcal{D}_f$,

$$\left|\widehat{f}_{n,k}(X_i) - f(X_i)\right| \leq \sqrt{\frac{\log(6n/\delta)}{2k}} + \omega\cdot\left(\frac{2k}{n}\right)^{\lambda}. \tag{8}$$

Given (8) we can upper bound $\widehat{M}_{n,k}(f)$ by

$$\widehat{M}_{n,k}(f) \leq \max_{i\in[n]}\{f(X_i)\} + \sqrt{\frac{\log(6n/\delta)}{2k}} + \omega\cdot\left(\frac{2k}{n}\right)^{\lambda}$$

$$\leq M(f) + \sqrt{\frac{\log(6n/\delta)}{2k}} + 2\omega\cdot\left(\frac{2k}{n}\right)^{\lambda}.$$

We now lower bound $\widehat{M}_{n,k}(f)$ as follows. Take $\epsilon > 0$ and choose $x_0 \in \mathcal{X}_\mu$ with $f(x_0) \geq M(f) - \epsilon$ (a point that nearly achieves the supremum of f). By Lemma 4.1 with probability at least $1-\delta/2$ over $\mathcal{D}_f$ we have

$$\left|f\left(X_{\tau_{n,1}(x_0)}\right) - f(x_0)\right| \leq \omega\cdot\left(\frac{2k}{n}\right)^{\lambda}. \tag{9}$$

In conjunction, the bounds (8) and (9) imply

$$\begin{aligned}\widehat{M}_{n,k}(f) &\geq \hat{f}_{n,k}\left(X_{\tau_{n,1}(x_0)}\right) \\ &\geq f\left(X_{\tau_{n,1}(x_0)}\right) - \sqrt{\frac{\log(6n/\delta)}{2k}} - \omega\cdot\left(\frac{2k}{n}\right)^{\lambda} \\ &\geq f(x_0) - \sqrt{\frac{\log(6n/\delta)}{2k}} - 2\omega\cdot\left(\frac{2k}{n}\right)^{\lambda} \\ &\geq M(f) - \sqrt{\frac{\log(6n/\delta)}{2k}} - 2\omega\cdot\left(\frac{2k}{n}\right)^{\lambda} - \epsilon.\end{aligned}$$

Combining this with the upper bound we see that given (8) and (9) we have

$$\left|\widehat{M}_{n,k}(f) - M(f)\right| \leq \sqrt{\frac{\log(6n/\delta)}{2k}} + 2\omega\cdot\left(\frac{2k}{n}\right)^{\lambda} + \epsilon. \tag{10}$$

By the union bound, the bounds (8) and (9), and hence (10) hold simultaneously with probability at least $1-\delta$. Letting $\epsilon \to 0$ and applying continuity of measure completes the proof of the theorem. □

4.3. Main result: Fast rates in the presence of unknown asymmetric label noise

We put everything together in this section, and complete the analysis of the label noise robust kNN classifier given in Algorithm 2. This classifier is simply the kNN instantiation of Algorithm 1. Moreover, this algorithm was previously proposed by (Gao et al., 2018) without analysis of its finite sample behaviour when the noise probabilities are unknown, and has been empirically demonstrated to be successful in practice. We shall now prove that it attains the known minimax optimal rates *of the noiseless setting*, up to a log factor, despite the presence of unknown asymmetric label noise, provided the assumptions discussed in Section 3.1.

Algorithm 2 A k nearest neighbour method for label noise

1. Define $\hat{\eta}_{\text{corr}}$ by $\hat{\eta}_{\text{corr}}(x) := \frac{1}{k}\cdot\sum_{q\in[k]}\tilde{Y}_{\tau_{n,q}(x)}$;
2. Compute $\hat{p}_0 := \min_{i\in[n]}\{\hat{\eta}_{\text{corr}}(X_i)\}$ and $\hat{p}_1 := 1 - \max_{i\in[n]}\{\hat{\eta}_{\text{corr}}(X_i)\}$;
3. Let $\hat{\phi}_{n,k}(x) := \mathbb{1}\{\hat{\eta}_{\text{corr}}(x) \geq 1/2\cdot(1+\hat{p}_0-\hat{p}_1)\}$.

Theorem 4.3. *Suppose that Assumptions 1 and 2 hold, Assumption 3 holds with exponent $\lambda > 0$ and constant $\omega > 0$, and Assumption 4 holds with constant $\alpha \geq 0$ and $C_\alpha \geq 1$. Take any $n \in \mathbb{N}$, $\delta \in (0,1)$, and suppose we have $\mathcal{D}_{corr} = \{(X_i, \tilde{Y}_i)\}_{i\in[n]}$ with $(X_i, \tilde{Y}_i) \sim \mathbb{P}_{corr}$. Let $\hat{\phi}_{n,k}$ be the label-noise robust kNN classifier with an arbitrary choice of $k \in \mathbb{N}\cap[4\log(3/\delta)+1, n/2]$ (Algorithm 2). With probability at least $1-\delta$ over $\mathcal{D}_{corr}$, we have*

$$\begin{aligned}\mathcal{R}\left(\hat{\phi}_{n,k}\right) \leq \mathcal{R}(\phi_*) + C_\alpha\cdot\left(\frac{8}{1-p_0-p_1}\right)^{\alpha+1} \\ \cdot\left[\sqrt{\frac{\log(18n/\delta)}{k}} + 2\omega\cdot\left(\frac{2k}{n}\right)^{\lambda}\right]^{\alpha+1} + \delta,\end{aligned} \tag{11}$$

where $\phi_*(x) \equiv \mathbb{1}\{\eta(x) \geq 1/2\}$ is the Bayes classifier.

In particular, if we take $k_n = \left\lceil\left(\frac{\log(18n/\delta)}{2\omega^2}\right)^{\frac{1}{2\lambda+1}}\cdot n^{\frac{2\lambda}{2\lambda+1}}\right\rceil$ then for $n \geq 5\cdot(10\cdot\omega^2)^{\frac{1}{2\lambda}}\cdot\log(18n/\delta)$, w.p. at least $1-\delta$,

$$\begin{aligned}\mathcal{R}\left(\hat{\phi}_{n,k_n^*}\right) \leq \mathcal{R}(\phi_*) + C_\alpha\cdot\left(\frac{2^{2\lambda+5}\cdot\omega^{\frac{1}{2\lambda+1}}}{1-p_0-p_1}\right)^{\alpha+1} \\ \cdot\left(\frac{\log(18n/\delta)}{n}\right)^{\frac{\lambda(\alpha+1)}{2\lambda+1}} + \delta.\end{aligned}$$

Proof. Observe first that the measure smoothness property of η implies that η_{corr} is measure smooth with the same exponent, and constant $(1-p_0-p_1)\cdot\omega \leq \omega$, so we can work with the latter to avoid clutter.

We define the subset of the input domain where the corrupted regression function has low estimation error:

$$\mathcal{G}_\delta := \{x \in \mathcal{X}_\mu : |\hat{\eta}_{\text{corr}}(x) - \eta_{\text{corr}}(x)| \leq \xi(n,k,\delta)\}.$$

where $\xi(n,k,\delta)$ is a small error that will be made precise shortly. We want to ensure that a randomly drawn test point is in this set w.p. $1-\delta/3$. By Theorem 4.1, for each $x \in \mathcal{X}_\mu$, the following holds with probability at least $1-\delta^2/3$,

$$\begin{aligned}|\hat{\eta}_{\text{corr}}(x) - \eta_{\text{corr}}(x)| &\leq \sqrt{\frac{\log(3/(\delta^2/3))}{2k}} + \omega\cdot\left(\frac{2k}{n}\right)^{\lambda} \\ &= \sqrt{\frac{2\log(3/\delta)}{2k}} + \omega\cdot\left(\frac{2k}{n}\right)^{\lambda} \\ &=: \xi_1(n,k,\delta).\end{aligned}$$

That is, for each fixed $x \in \mathcal{X}_\mu$, we have $x \in \mathcal{G}_\delta$ with probability at least $1-\delta^2/3$ i.e. $\mathbb{E}_{\mathcal{D}_f}[\mathbb{1}\{x \notin \mathcal{G}_\delta\}] \leq \delta^2/3$. We now integrate over μ and use Fubini's theorem as follows:

$$\begin{aligned}\mathbb{E}_{\mathcal{D}_f}\left[\mu\left(\mathcal{X}_\mu\backslash\mathcal{G}_\delta\right)\right] &= \mathbb{E}_{\mathcal{D}_f}\left[\int\mathbb{1}\{x\notin\mathcal{G}_\delta\}\,d\mu(x)\right] \\ &= \int\mathbb{E}_{\mathcal{D}_f}\left[\mathbb{1}\{x\notin\mathcal{G}_\delta\}\right]d\mu(x) \leq \delta^2/3.\end{aligned}$$

Thus, by Markov's inequality we have $\mu\left(\mathcal{X}_\mu\backslash\mathcal{G}_\delta\right) \leq \delta$ with probability at least $1-\delta/3$. Furthermore, by Theorem 4.2 with probability at least $1-\delta/3$,

$$|\hat{p}_0 - p_0| \leq \sqrt{\frac{\log(6n/(\delta/3))}{2k}} + 2\omega\cdot\left(\frac{2k}{n}\right)^{\lambda} =: \xi_2(n,k,\delta).$$

Similarly, with probability at least $1-\delta/3$ we have $|\hat{p}_1 - p_1| \leq \xi_2(n,k,\delta)$, and we let

$$\begin{aligned}\xi(n,k,\delta) &:= \max\{\xi_1(n,k,\delta), \xi_2(n,k,\delta)\} \\ &\leq \sqrt{\frac{\log(18n/\delta)}{k}} + 2\omega \cdot \left(\frac{2k}{n}\right)^{\lambda}. \qquad (12)\end{aligned}$$

By the union bound, with probability at least $1-\delta$, we have $\mu\left(\mathcal{X}_\mu \backslash \mathcal{G}_\delta\right) \leq \delta$ and $\max\left\{|\hat{p}_0 - p_0|, |\hat{p}_1 - p_1|\right\} \leq \xi(n,k,\delta)$. Hence, it suffices to assume that $\mu\left(\mathcal{X}_\mu \backslash \mathcal{G}_\delta\right) \leq \delta$ and $\max\left\{|\hat{p}_0 - p_0|, |\hat{p}_1 - p_1|\right\} \leq \xi(n,k,\delta)$ holds and show that eq (11) holds.

Observe that we can rewrite $\hat{\phi}_n : \mathcal{X} \to \mathcal{Y}$ as $\hat{\phi}_n(x) = \mathbb{1}\left\{\hat{\eta}(x) \geq 1/2\right\}$, where $\hat{\eta}(x) := \left(\hat{\eta}_{\text{corr}}(x) - \hat{p}_0\right) / \left(1 - \hat{p}_0 - \hat{p}_1\right)$. By Lemma 3.1 for all $x \in \mathcal{G}_\delta$ we have deterministically that:

$$\begin{aligned}&|\hat{\eta}(x) - \eta(x)| \\ &\leq 8 \cdot \frac{\max\left\{|\hat{\eta}_{\text{corr}}(x) - \eta_{\text{corr}}(x)|, |\hat{p}_0 - p_0|, |\hat{p}_1 - p_1|\right\}}{(1 - p_0 - p_1)} \\ &\leq 8 \cdot (1 - p_0 - p_1)^{-1} \cdot \xi(n,k,\delta).\end{aligned}$$

Hence, observe that, given any $x \in \mathcal{X}$ with $\hat{\phi}_n(x) \neq \phi_*(x) \equiv \mathbb{1}\left\{\eta(x) \geq 1/2\right\}$ we must have $|\eta(x) - 1/2| \leq 8 \cdot (1 - p_0 - p_1)^{-1} \cdot \xi(n,k,\delta)$. Hence, by Assumption 4, with probability at least $1-\delta$, we have

$$\begin{aligned}&\mathcal{R}\left(\hat{\phi}_{n,k}\right) - \mathcal{R}\left(\phi_*\right) \\ &= \int_{\mathcal{X}} \left|\eta(x) - \frac{1}{2}\right| \cdot \mathbb{1}\left\{\hat{\phi}_{n,\delta}(x) \neq \phi_*(x)\right\} d\mu(x) \\ &\leq \int_{\mathcal{G}_\delta} \left|\eta(x) - \frac{1}{2}\right| \cdot \mathbb{1}\left\{\hat{\phi}_{n,\delta}(x) \neq \phi_*(x)\right\} d\mu(x) + \mu\left(\mathcal{X} \backslash \mathcal{G}_\delta\right) \\ &\leq \int_{\mathcal{X}} \left|\eta(x) - \frac{1}{2}\right| \cdot \mathbb{1}\left\{\left|\eta(x) - \frac{1}{2}\right| \leq \frac{8 \cdot \xi(n,k,\delta)}{1 - p_0 - p_1}\right\} d\mu(x) + \delta \\ &\leq C_\alpha \cdot \left(\frac{8 \cdot \xi(n,k,\delta)}{1 - p_0 - p_1}\right)^{\alpha+1} + \delta.\end{aligned}$$

Plugging in eq. (12) completes the proof of the first part.

The second part follows by choosing k that approximately equates the two terms on the right hand side of eq. (12). That is, with the choice

$$k_n = \left\lceil \left(\frac{\log(18n/\delta)}{2\omega^2}\right)^{\frac{1}{2\lambda+1}} \cdot n^{\frac{2\lambda}{2\lambda+1}} \right\rceil,$$

given $n \geq 5 \cdot (10 \cdot \omega^2)^{\frac{1}{2\lambda}} \cdot \log(18n/\delta)$ we have $k_n \geq 4\log(3/\delta) + 1$ and $\xi(n,k_n,\delta)$ takes the form

$$\xi(n,k_n,\delta) = 4^{\lambda+1} \cdot \omega^{\frac{1}{2\lambda+1}} \cdot \left(\frac{\log(18n/\delta)}{n}\right)^{\frac{\lambda}{2\lambda+1}}.$$

Plugging this into the excess risk completes the proof of the second part of the theorem. □

4.3.1. On setting the value of k

We used the theoretically optimal value of k in our analysis, which of course is not available in practice. Methods exist to set k in a data-driven manner. Cross-validation is amongst the most popular practical approaches (Inouye et al., 2017; Gao et al., 2018), and there is also an ample literature on adaptive non-parametric estimation methods (e.g. Chapter 8, (Giné & Nickl, 2015)) that are not known to be practical but allow us to retain nearly optimal rates without access to the unknown parameters of the analysis.

5. Discussion: Inconsistency of kNN in the presence of asymmetric label noise

Our main result implies that under the measure-smoothness condition, and provided the label-noise identifiability conditions hold, the statistical difficulty of classification with or without asymmetric label noise is the same in the minimax sense, up to constants and log factors. However, the algorithm that we used to achieve this rate was not the classical kNN. This invites the question as to whether this must be so (or is it an artefact of the proof)? We find this question interesting in the light of observations and claims in the literature about the label-noise robustness of kNN and other high capacity models (Tarlow et al., 2013; Gao et al., 2018) (see also the introductory section of (Menon et al., 2018)).

To shed light on this, we show in this section that the classical kNN cannot achieve these rates, and even fails to be consistent in the presence of asymmetric label noise. In fact, in Theorem 5.1 below we shall see that *any* algorithm that is Bayes-consistent in the classical sense may become inconsistent under this type of noise. Indeed, on closer inspection, the robustness claims in the literature about kNN and other high capacity models (Tarlow et al., 2013; Gao et al., 2018; Menon et al., 2018) – explicitly or tacitly – refer either to class-unconditional (i.e. symmetric) label noise, or assume that the regression function is bounded away from $1/2$. A recent result of (Cannings et al., 2018) even gives convergence rates for k-NN under instance-dependent label noise, but requires that the label noise probabilities become symmetric as the regression function approaches $1/2$.

In order to talk about consistency in the presence of label noise we need to make explicit the distinct roles of the train and test distributions in our notation. For any distribution $\mathbb{P}$ on $\mathcal{X} \times \{0,1\}$ and classifier $\phi : \mathcal{X} \to \{0,1\}$ the risk is defined as $\mathcal{R}\left(\phi\right) = \mathcal{R}\left(\phi; \mathbb{P}\right) := \mathbb{E}\left[\phi(X) \neq Y\right]$, where $\mathbb{E}$ denotes the expectation with respect to $\mathbb{P}$. In addition we let $\mathcal{E}\left(\phi; \mathbb{P}\right)$ denote the excess risk defined by $\mathcal{E}\left(\phi, \mathbb{P}\right) := \mathcal{R}(\phi; \mathbb{P}) - \inf_{\tilde{\phi}}\{\mathcal{R}(\tilde{\phi}; \mathbb{P})\}$, where the infimum is over all measurable functions $\tilde{\phi} : \mathcal{X} \to \{0,1\}$.

Definition 5.1 (Consistency)**.** *Let $\mathbb{P}_{train}$,$\mathbb{P}_{test}$ be probability distributions on $\mathcal{X} \times \{0,1\}$. Take $\tilde{\mathcal{D}} = \{(X_i, \tilde{Y}_i)\}_{i \in \mathbb{N}}$*

with $(X_i, \tilde{Y}_i)$ sampled independently from $\mathbb{P}_{train}$. For each $n \in \mathbb{N}$ we let $\hat{\phi}_n$ denote the classifier obtained by applying the learning algorithm $\hat{\phi}$ to the data set $\tilde{\mathcal{D}}$. We shall say that a classification algorithm $\hat{\phi}$ is consistent with training distribution $\mathbb{P}_{train}$ and test distribution $\mathbb{P}_{test}$ if we have $\lim_{n\to\infty} \mathcal{E}(\hat{\phi}_n; \mathbb{P}_{test}) = 0$ almost surely over the data $\tilde{\mathcal{D}}$.

Let $\mathbb{P}_{(\mu,\eta)}$ denote the probability distribution on $\mathcal{X} \times \{0,1\}$ with marginal μ and a regression function η. Learning with label noise means that the training distribution $\mathbb{P}_{(\mu,\eta_{\text{corr}})}$ and test distribution $\mathbb{P}_{(\mu,\eta)}$ are different. We define the input set of disagreement:

$$\mathcal{A}_0(\eta, \eta_{\text{corr}}) := \left\{ x \in \mathcal{X} : \left(\eta(x) - \frac{1}{2}\right)\left(\eta_{\text{corr}}(x) - \frac{1}{2}\right) < 0 \right\}.$$

Note that $\mathcal{A}_0(\eta, \eta_{\text{corr}})$ consists of points which are sufficiently close to the boundary with asymmetric label noise.

Theorem 5.1. *Suppose that a classification algorithm $\hat{\phi}$ is consistent with $\mathbb{P}_{train} = \mathbb{P}_{test} = \mathbb{P}_{(\mu,\eta_{corr})}$, and let $\eta \neq \eta_{corr}$. If $\mu\left(\mathcal{A}_0(\eta, \eta_{corr})\right) > 0$ then $\hat{\phi}$ is inconsistent with $\mathbb{P}_{train} = \mathbb{P}_{(\mu,\eta_{corr})}$ and $\mathbb{P}_{test} = \mathbb{P}_{(\mu,\eta)}$.*

The proof is given in Appendix. The essence of the argument is the simple observation that if the regression functions of the training and testing distributions disagree, then the trained classifier cannot agree with both. Below we give a family of examples on $\mathcal{X} = \mathbb{R}$ with class-conditional label noise, where the standard k_n-NN classifier is inconsistent, yet the k_n-NN method for asymmetric label noise (Algorithm 2) is consistent.

Example: Take any $p_0, p_1 \in (0, 1/2)$ with $p_0 \neq p_1$ and let $m := (2 - 3p_0 - p_1) / (4\,(1 - p_0 - p_1))$. It follows that $m \in (0,1)$. Let $\mathcal{X} = [0,1]$ and let μ be the Lebesgue measure on $\mathcal{X}$. Define $\eta : \mathcal{X} \to [0,1]$ by

$$\eta(x) := \begin{cases} \frac{3x}{2} & \text{if } x \in \left[0, \frac{2m}{3}\right] \\ m & \text{if } x \in \left[\frac{2m}{3}, \frac{2m+1}{3}\right] \\ \frac{3x-1}{2} & \text{if } x \in \left[\frac{2m+1}{3}, 1\right]. \end{cases}$$

A special case of this example, with $p_0 = 0.1$ and $p_1 = 0.3$ is depicted in Figure 1.

Indeed, for this family of examples, it follows from Theorem 1 in (Chaudhuri & Dasgupta, 2014) that the standard k_n-NN classifier is strongly Bayes-consistent with $\mathbb{P}_{train} = \mathbb{P}_{test} = \mathbb{P}_{(\mu,\eta_{\text{corr}})}$ whenever $k_n/n \to 0$ and $k_n/(\log(n)) \to \infty$ as $n \to \infty$. Moreover, it follows from the definition of m, $p_0 \neq p_1$, and eq. (1) that for $x \in \left[\frac{2m}{3}, \frac{2m+1}{3}\right]$ we have

$$\left(\eta(x) - \frac{1}{2}\right)\left(\eta_{\text{corr}}(x) - \frac{1}{2}\right) < 0.$$

Hence, $\mu\left(\mathcal{A}_0(\eta, \eta_{\text{corr}})\right) = 1/3 > 0$. Thus, by Theorem 5.1, the k_n-NN classifier is inconsistent with $\mathbb{P}_{train} = \mathbb{P}_{(\mu,\eta_{\text{corr}})}$ and $\mathbb{P}_{test} = \mathbb{P}_{(\mu,\eta)}$. On the other hand, one can readily check that Assumptions 2 and 1 hold. Moreover, Assumption 3 holds with exponent $\lambda = 1$ and constant $\omega = 3$, and Assumption 4 holds with exponent $\alpha = 0$ and constant $C_\alpha = 1$. Thus, by Theorem 4.3 combined with the Borel-Cantelli lemma, the k_n method for asymmetric label noise (Algorithm 2) is consistent with $\mathbb{P}_{train} = \mathbb{P}_{(\mu,\eta_{\text{corr}})}$ and $\mathbb{P}_{test} = \mathbb{P}_{(\mu,\eta)}$ whenever $k_n/n \to 0$ and when $k_n/(\log(n)) \to \infty$ as $n \to \infty$.

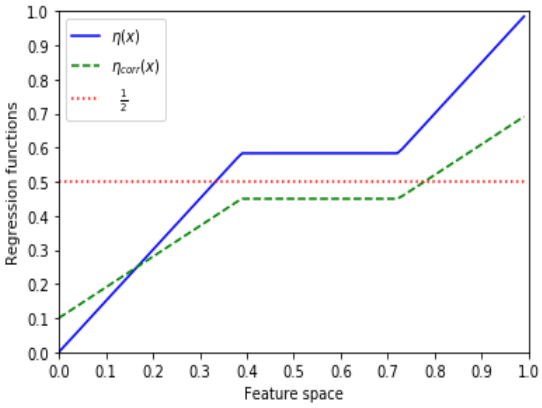

Figure 1. A pair of train and test regression functions $(\eta, \eta_{\text{corr}})$ exemplifying a setting where classical kNN is inconsistent yet the kNN method for asymmetric label noise is consistent.

6. Conclusions

We obtained fast rates in the presence of unknown asymmetric label noise that match the minimax optimal rates of the noiseless setting, up to a log factor, under measure smoothness and Tsybakov margin assumptions. On the practical side, our results provide theoretical support for the Robust kNN algorithm of (Gao et al., 2018) whose analysis so far only exists under known noise probabilities. On the theoretical side, our results entail that under the stated conditions the statistical difficulty of classification with or without unknown asymmetric label noise is the same in the minimax sense. This is especially interesting given recent results which show that under more general non-parametric settings the optimal rates for unknown asymmetric label noise can be strictly slower than the those for the noiseless case (Reeve & Kaban, 2019). We have also seen that the algorithm achieving the rate in the presence of unknown asymmetric label noise must be different from any classical Bayes-consistent classifier, as those fail to be consistent under the label noise. Finally, a key ingredient in our analysis a simple method for estimating the maximum of a function that requires far less assumptions than existing mode estimators do and may have wider applicability.

Acknowledgement

This work is funded by EPSRC under Fellowship grant EP/P004245/1, and a Turing Fellowship (grant EP/N510129/1). We would also like to thank the anonymous reviewers for their careful feedback.

References

Audibert, J.-Y., Tsybakov, A. B., et al. Fast learning rates for plug-in classifiers. *The Annals of statistics*, 35(2): 608–633, 2007.

Biau, G. and Devroye, L. *Lectures on the Nearest Neighbor Method*. Springer Publishing Company, Incorporated, 1st edition, 2015.

Blanchard, G., Lee, G., and Scott, C. Semi-supervised novelty detection. *Journal of Machine Learning Research*, 11(Nov):2973–3009, 2010.

Blanchard, G., Flaska, M., Handy, G., Pozzi, S., and Scott, C. Classification with asymmetric label noise: Consistency and maximal denoising. *Electron. J. Statist.*, 10(2): 2780–2824, 2016.

Boucheron, S., Lugosi, G., and Massart, P. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

Cannings, T. I., Fan, Y., and Samworth, R. J. Classification with imperfect training labels. *ArXiv e-prints*, May 2018.

Chaudhuri, K. and Dasgupta, S. Rates of convergence for nearest neighbor classification. In *Advances in Neural Information Processing Systems*, pp. 3437–3445, 2014.

Dasgupta, S. and Kpotufe, S. Optimal rates for k-nn density and mode estimation. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, pp. 2555–2563, Cambridge, MA, USA, 2014. MIT Press.

Döring, M., Györfi, L., and Walk, H. Rate of convergence of k-nearest-neighbor classification rule. *Journal of Machine Learning Research*, 18:227:1–227:16, 2017.

Elkan, C. and Noto, K. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, New York, NY, USA, 2008. ACM.

Frénay, B. and Verleysen, M. Classification in the presence of label noise: A survey. *IEEE Trans. Neural Netw. Learning Syst.*, 25(5):845–869, 2014.

Gao, W., Niu, X., and Zhou, Z. On the consistency of exact and approximate nearest neighbor with noisy data. *Arxiv*, abs/1607.07526, 2018.

Giné, E. and Nickl, R. *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2015.

Inouye, D. I., Ravikumar, P., Das, P., and Dutta, A. Hyperparameter selection under localized label noise via corrupt validation. In *NIPS Workshop*. 2017.

Jiang, H. Non-asymptotic uniform rates of consistency for k-nn regression. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*. AAAI, 2019.

Jiang, H. and Kpotufe, S. Modal-set estimation with an application to clustering. In *Artificial Intelligence and Statistics*, pp. 1197–1206, 2017.

Kpotufe, S. k-nn regression adapts to local intrinsic dimension. In Shawe-Taylor, J., Zemel, R. S., Bartlett, P. L., Pereira, F., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems 24*, pp. 729–737. 2011.

Mammen, E. and Tsybakov, A. B. Smooth discrimination analysis. *Ann. Statist.*, 27(6):1808–1829, 12 1999.

Menon, A., Van Rooyen, B., Ong, C. S., and Williamson, B. Learning from corrupted binary labels via class-probability estimation. In *International Conference on Machine Learning*, pp. 125–134, 2015.

Menon, A. K., van Rooyen, B., and Natarajan, N. Learning from binary labels with instance-dependent noise. *Machine Learning*, 107(8):1561–1595, Sep 2018.

Mitzenmacher, M. and Upfal, E. *Probability and computing: Randomized algorithms and probabilistic analysis*. Cambridge university press, 2005.

Natarajan, N., Dhillon, I. S., Ravikumar, P. K., and Tewari, A. Learning with noisy labels. In *Advances in neural information processing systems*, pp. 1196–1204, 2013.

Natarajan, N., Dhillon, I. S., Ravikumar, P., and Tewari, A. Cost-sensitive learning with noisy labels. *Journal of Machine Learning Research*, 18(155):1–33, 2018.

Reeve, H. W. and Kaban, A. Classification with unknown class conditional label noise on non-compact feature spaces. *Accepted to COLT*, 2019.

Scott, C. A rate of convergence for mixture proportion estimation, with application to learning from noisy labels. In *Artificial Intelligence and Statistics*, pp. 838–846, 2015.

Scott, C., Blanchard, G., and Handy, G. Classification with asymmetric label noise: Consistency and maximal denoising. In *Conference On Learning Theory*, pp. 489–511, 2013.

Tarlow, D., Swersky, K., Charlin, L., Sutskever, I., and Zemel, R. Stochastic k-neighborhood selection for supervised and unsupervised learning. In Dasgupta, S. and McAllester, D. (eds.), *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 199–207, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR.

Tsybakov, A. B. *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 1st edition, 2008.

A. Proof of Theorem 5.1

The proof of Theorem 5.1 is as follows.

Proof. Define the input set of disagreement with margin θ:

$$\mathcal{A}_\theta(\eta, \eta_{\text{corr}}) := \left(\left\{\eta(x) \leq \frac{1}{2} - \theta\right\} \cap \left\{\eta_{\text{corr}}(x) \geq \frac{1}{2} + \theta\right\}\right) \tag{13}$$

$$\cup \left(\left\{\eta(x) \geq \frac{1}{2} + \theta\right\} \cap \left\{\eta_{\text{corr}}(x) \leq \frac{1}{2} - \theta\right\}\right). \tag{14}$$

We can write $\mathcal{A}_0(\eta, \eta_{\text{corr}})$ as a union of such sets: $\mathcal{A}_0(\eta) = \bigcup_{\theta>0} \mathcal{A}_\theta(\eta)$, and hence

$$\lim_{\theta \to 0} \{\mu(\mathcal{A}_\theta(\eta, \eta_{\text{corr}}))\} = \mu(\mathcal{A}_0(\eta, \eta_{\text{corr}})) > 0.$$

Now, take some $\theta > 0$ s.t. $\mathcal{A}_0(\eta, \eta_{\text{corr}}) > 0$. Lemma A.1 below will show that

$$\mathcal{E}\left(\phi; \mathbb{P}_{(\mu,\eta)}\right) + \mathcal{E}\left(\phi; \mathbb{P}_{(\mu,\eta_{\text{corr}})}\right) \geq \theta \cdot \mu(\mathcal{A}_\theta(\eta, \eta_{\text{corr}})) > 0. \tag{15}$$

Since $\hat{\phi}$ is consistent with $\mathbb{P}_{train} = \mathbb{P}_{test} = \mathbb{P}_{(\mu,\eta_{\text{corr}})}$, we have $\lim_{n\to\infty} \mathcal{E}\left(\hat{\phi}_n; \mathbb{P}_{(\mu,\eta_{\text{corr}})}\right) = 0$. Hence, from eq. (15) it follows that $\limsup_{n\to\infty} \mathcal{E}\left(\hat{\phi}_n; \mathbb{P}_{(\mu,\eta)}\right) \geq \theta \cdot \mu(\mathcal{A}_\theta(\eta, \eta_{\text{corr}})) > 0$. That is, $\hat{\phi}$ is inconsistent when trained with train distribution $\mathbb{P}_{(\mu,\eta_{\text{corr}})}$ and tested on distribution $\mathbb{P}_{(\mu,\eta)}$. □

It remains to prove the lemma used in the proof above.

Lemma A.1. *Let μ be a Borel probability measure on $\mathcal{X}$. Given $\eta : \mathcal{X} \to [0,1]$ and $\theta > 0$ consider the set $\mathcal{A}_\theta(\eta, \eta_{corr}) \subseteq \mathcal{X}$ as defined in eq. (14). Then given any classifier $\phi : \mathcal{X} \to \{0,1\}$ we have $\mathcal{E}\left(\phi; \mathbb{P}_{(\mu,\eta)}\right) + \mathcal{E}\left(\phi; \mathbb{P}_{(\mu,\eta_{corr})}\right) \geq \theta \cdot \mu(\mathcal{A}_\theta(\eta, \eta_{corr}))$.*

Proof. Recall that, for any regression function $\tilde{\eta} : \mathcal{X} \to [0,1]$ the excess risk can be written as: $\mathcal{E}\left(\phi; \mathbb{P}_{(\mu,\tilde{\eta})}\right) =$

$$\int \left|\tilde{\eta}(x) - \frac{1}{2}\right| \cdot \mathbb{1}\left\{\left(\tilde{\eta}(x) - \frac{1}{2}\right)\left(\phi(x) - \frac{1}{2}\right) < 0\right\} d\mu(x). \tag{16}$$

Now if $x \in \mathcal{A}_\theta(\eta, \eta_{\text{corr}})$ then $\left(\eta(x) - \frac{1}{2}\right)\left(\eta_{\text{corr}}(x) - \frac{1}{2}\right) < 0$ so for both possible values $\phi(x) \in \{0,1\}$ we have

$$\mathbb{1}\left\{\left(\eta(x) - \frac{1}{2}\right)\left(\phi(x) - \frac{1}{2}\right) < 0\right\} + \mathbb{1}\left\{\left(\eta_{\text{corr}}(x) - \frac{1}{2}\right)\left(\phi(x) - \frac{1}{2}\right) < 0\right\} = 1.$$

Moreover, if $x \in \mathcal{A}_\theta(\eta, \eta_{\text{corr}})$ then $\min\left\{\left|\eta(x) - \frac{1}{2}\right|, \left|\eta_{\text{corr}}(x) - \frac{1}{2}\right|\right\} \geq \theta$ and so

$$\left|\eta(x) - \frac{1}{2}\right| \cdot \mathbb{1}\left\{\left(\eta(x) - \frac{1}{2}\right)\left(\phi(x) - \frac{1}{2}\right) < 0\right\} + \left|\eta_{\text{corr}}(x) - \frac{1}{2}\right| \cdot \mathbb{1}\left\{\left(\eta_{\text{corr}}(x) - \frac{1}{2}\right)\left(\phi(x) - \frac{1}{2}\right) < 0\right\} \geq \theta. \tag{17}$$

Integrating with respect to μ and applying (16) to both $\mathbb{P}_{(\mu,\eta)}$ and $\mathbb{P}_{(\mu,\eta_{\text{corr}})}$ gives the conclusion of the lemma. □

B. Proof of Lemma 3.1

Proof. Given $\hat{a}, a \in [-1,1]$, $b, \hat{b} > 0$ with $|\hat{b} - b| \leq b/2$, and $a/b \in [0,1]$,

$$\left|\frac{\hat{a}}{\hat{b}} - \frac{a}{b}\right| = \frac{1}{\hat{b}} \cdot \left|(\hat{a} - a) + \frac{a}{b} \cdot (b - \hat{b})\right| \leq \frac{2}{b}\left(|\hat{a} - a| + \left|\frac{a}{b}\right| \cdot |\hat{b} - b|\right) \leq \frac{4}{b} \cdot \max\left\{|\hat{a} - a|, |\hat{b} - b|\right\}, \tag{18}$$

where we have used the fact that $\hat{b} \geq b/2$. By the definition of $\hat{\eta}(x)$ together with eq. (1) we have

$$\hat{\eta}(x) := \frac{\hat{\eta}_{\text{corr}}(x) - \hat{p}_0}{1 - \hat{p}_0 - \hat{p}_1} \qquad \text{and} \qquad \eta(x) = \frac{\eta_{\text{corr}}(x) - p_0}{1 - p_0 - p_1}.$$

Now take $\hat{a} = \hat{\eta}_{\text{corr}}(x) - \hat{p}_0$, $a = \eta_{\text{corr}}(x) - p_0$, $\hat{b} = 1 - \hat{p}_0 - \hat{p}_1$ and $b = 1 - p_0 - p_1$. Given the assumptions that $p_0 + p_1 < 1$, so $b > 0$ and $\max\{|\hat{p}_0 - p_0|, |\hat{p}_1 - p_1|\} \leq (1 - p_0 - p_1)/4$ this implies

$$|\hat{b} - b| = 2 \cdot \max\{|\hat{p}_0 - p_0|, |\hat{p}_1 - p_1|\} \leq \frac{1}{2} \cdot (1 - p_0 - p_1) = \frac{b}{2},$$

which also implies $\hat{b} \geq b/2 > 0$. Hence, by (18) we deduce

$$\begin{aligned}|\hat{\eta}(x) - \eta(x)| &\leq \frac{4}{1 - p_0 - p_1} \cdot \max\{|(\hat{\eta}_{\text{corr}}(x) - \hat{p}_0) - (\eta_{\text{corr}}(x) - p_0)|, |(1 - \hat{p}_0 - \hat{p}_1) - (1 - p_0 - p_1)|\} \\ &\leq \frac{8}{1 - p_0 - p_1} \cdot \max\{|\hat{\eta}_{\text{corr}}(x) - \eta_{\text{corr}}(x)|, |\hat{p}_0 - p_0|, |\hat{p}_1 - p_1|\}.\end{aligned}$$

This completes the proof of Lemma 3.1. □