Hypothesis Testing for Class-Conditional Label Noise

Rafael Poyiadzi✉, Weisong Yang, Niall Twomey, and Raul Santos-Rodriguez

University of Bristol, UK
{rp13102,ws.yang,niall.twomey,enrsr}@bristol.ac.uk

Abstract. In this paper we aim to provide machine learning practitioners with tools to answer the question: *have the labels in a dataset been corrupted?* In order to simplify the problem, we assume the practitioner already has preconceptions on possible distortions that may have affected the labels, which allow us to pose the task as the design of hypothesis tests. As a first approach, we focus on scenarios where a given dataset of instance-label pairs has been corrupted with *class-conditional label noise*, as opposed to *uniform label noise*, with the former biasing learning, while the latter – under mild conditions – does not. While previous works explore the direct estimation of the noise rates, this is known to be hard in practice and does not offer a real understanding of how trustworthy the estimates are. These methods typically require *anchor points* – examples whose true posterior is either 0 or 1. Differently, in this paper we assume we have access to a set of anchor points whose true posterior is approximately 1/2. The proposed hypothesis tests are built upon the asymptotic properties of Maximum Likelihood Estimators for Logistic Regression models. We establish the main properties of the tests, including a theoretical and empirical analysis of the dependence of the power on the test on the training sample size, the number of anchor points, the difference of the noise rates and the use of relaxed anchors.

1 Introduction

When a machine learning practitioner is presented with a new dataset, a first question is that of data quality ([24]) as this will affect any subsequent machine learning tasks. This has led to tools to address transparency and accountability of data ([27,28]). However, in supervised learning, an equally important concern is the quality of labels. For instance, in standard data collections, data curators usually rely on annotators from online platforms, where individual annotators cannot be unconditionally trusted as they have been shown to perform inconsistently [25]. Labels are also expected to not be ideal in situations where the data is harvested directly from the web [31, 32]. In general this is a consequence of annotations not being carried out by domain experts [13].

The existing literature primarily focuses on directly estimating the distortion(s) present in the labels and mainly during the learning process (see Section 4). In this paper we argue that, in most cases, that is too hard a problem and might lead to suboptimal outcomes. Instead, we suggest modifying this approach in two ways. First, we leverage the practitioner's prior knowledge on the type possible distortions affecting the labels and use their preconceptions to design hypothesis testing procedures that would allow us (under certain assumptions we state later) to provide a measure of evidence for the presence of the distortion. This is of course a much simple task than addressing the estimation of any possible distortion. As an example, in this paper, we focus on class-conditional noise, as opposed to uniform noise (as we discuss later, class-conditional noise biases the learning procedure, while uniform noise under mild conditions does not). Secondly, with this information at hand, and given that the tests are performed right after data collection and annotation and before learning takes place, the practitioner can then make more informed decisions. If the quality of the labels is deemed poor, then the practitioner could resort to: (1) a modified data labelling procedure (e.g., active learning in the presence of noise [29]), (2) seek methods to make the training robust (e.g., algorithms for learning from noisy labels [30]), or (3) drop the dataset altogether.

Let us introduce the binary classification setting, where the goal is to train a classifier $g : \mathcal{X} \to \{-1, +1\}$, from a labelled dataset $\mathcal{D}_n^{train} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n \in (\mathbb{R}^d \times \{-1, 1\})$, with the objective of achieving a low miss-classification error: $\mathbb{P}_{X,Y}(g(X) \neq Y)$. While it is generally assumed that the training dataset is drawn from the distribution for which we wish to minimise the error for $\mathcal{D}_n^{train} \sim p(X, Y)$, as mentioned above, this is often not the case. Instead, the task requires us to train a classifier on a corrupted version of the dataset $\tilde{\mathcal{D}}_n^{train} \sim p(X, \tilde{Y})$ whilst still hoping to achieve a low error rate on the clean distribution.

In this work we focus on a particular type of corruption, *instance-independent label noise*, where labels are flipped with a certain rate, that can either be uniform across the entire data-generating distribution or conditioned on the true class of the data point. A motivating example of class-conditional noise is given in [12] in the form of medical case-control studies, where different tests may be used for subject and control. An essential ingredient in our procedure is the input from the user in the form of a set of *anchor points*. Differently from previous works, we assume anchor points for which the true posterior distribution $\mathbb{P}(Y = 1 \mid X = x)$ is (approximately) $1/2$. For an instance $\boldsymbol{x}$ this requirement means that an expert would not be able to provide *any* help to identify the correct class label. While this will be shown to be convenient for theoretical purposes, finding such anchor points might be rather difficult to accomplish in practice, so we show how to relax this notion to a more realistic $\eta(x) \approx 1/2$.

The tests rely on the asymptotic properties of the *Maximum Likelihood Estimate* (MLE) solution for Logistic Regression models, and the relationship between the true and noisy posteriors. On the theoretical side, we show that when the asymptotic properties of MLE hold and the user provides a single anchor point, we can devise hypothesis tests to assess the presence of class-conditional label corruption in the dataset. We then further extend these ideas to allow for richer sets of anchor points and illustrate how these lead to gains in the *power* of the test. In Section 2 we cover the necessary background on MLE, noisy labels and define the necessary tools. In Section 3 we illustrate how to carry a z-test using anchor points on the presence of class-conditional noise. In Section 4 we discuss related work and in Section 5 we present experimental findings.

2 Background

We are provided with a dataset $(\boldsymbol{X}, \boldsymbol{y}) = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n} \in (\mathbb{R}^d \times \{-1, 1\})$, and our task is to assess whether the labels have been corrupted with class-conditional flipping noise. We use y to denote the true label, and $\tilde{y}$ to denote the noisy label. We assume the feature vectors $(\boldsymbol{x})$ have been augmented with *ones* such that we have $\boldsymbol{x} \to (1,\ \boldsymbol{x})$. We assume the following model:

$$y_i \sim \text{Bernoulli}\left(\eta_i\right),$$
$$\eta_i = \sigma(\theta_0^\top \boldsymbol{x}_i) = \frac{1}{1 + \exp\left(-\theta_0^\top \boldsymbol{x}_i\right)}.$$

Following the MLE procedure we have:

$$\hat{\theta}_n = \underset{\theta \in \Theta}{\text{argmax}}\ \ell_n\left(\theta \mid D_n\right) = \underset{\theta \in \Theta}{\text{argmax}} \prod_{i=1}^{n} \ell_i\left(\theta \mid \boldsymbol{x}_i,\ y_i\right)$$

where:

$$\ell\left(\theta \mid \boldsymbol{x}_i,\ y_i\right) = \frac{y_i + 1}{2} \cdot \log \eta_i + \frac{1 - y_i}{2} \cdot \log(1 - \eta_i)$$

In this setting, the following can be shown (See for example Chapter 4 of [15]):

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \xrightarrow{D} \mathcal{N}\left(0,\ I_n(\theta_0)^{-1}\right) \tag{1}$$

where I_{θ_0} denotes the Fisher-Information Matrix:

$$I_n(\theta_0) = \mathbb{E}_\theta\left(-\frac{\partial^2 \ell_n(\theta; Y \mid x)}{\partial\theta\partial\theta^\top}\right) = \mathbb{E}_\theta\left(-H_n(\theta; Y \mid x)\right)$$

where the expectation is with respect to the conditional distribution, and H_n is the Hessian matrix.

We will consider two types of flipping noise and in both cases the noise rates are independent of the instance: $\mathbb{P}(\tilde{Y} = -i \mid Y = i,\ X = x) = \mathbb{P}(\tilde{Y} = -i \mid Y = i)$ for $i \in \{-1,\ 1\}$.

Definition 1. Bounded Uniform Noise (UN)
In this setting the per-class noise rates are identical: $\mathbb{P}(\tilde{Y} = 1 \mid Y = -1) = \mathbb{P}(\tilde{Y} = -1 \mid Y = 1) = \tau$ *and bounded:* $\tau < 0.50$*. We will denote this setting with UN(*τ*), and a dataset* $\mathcal{D} = (\boldsymbol{X}, \boldsymbol{y})$ *inflicted by UN(*τ*) by:* $\mathcal{D}_\tau$*.*

Definition 2. Bounded Class-Conditional Noise (CCN)
In this setting the per-class noise rates are different, $\alpha \neq \beta$ *and bounded* $\alpha + \beta < 1$ *with:* $\mathbb{P}(\tilde{Y} = -1 \mid Y = 1) = \alpha$ *and* $\mathbb{P}(\tilde{Y} = 1 \mid Y = -1) = \beta$*. We will denote this setting with CCN(*α, β*), and a dataset* $\mathcal{D} = (\boldsymbol{X}, \boldsymbol{y})$ *inflicted by CCN(*α, β*) by:* $\mathcal{D}_{\alpha,\beta}$*.*

An object of central interest in classification settings is the posterior predictive distribution: $\eta(\boldsymbol{x}) = \mathbb{P}(Y = 1 \mid X = \boldsymbol{x})$. Its noisy counterpart, $\tilde{\eta}(\boldsymbol{x}) = \mathbb{P}(\tilde{Y} = 1 \mid X = \boldsymbol{x})$, under the two settings, $UN(\tau)$ and $CCN(\alpha, \beta)$, can be expressed as: (See Appendix 8.1 for full derivation)

$$\tilde{\eta}(\boldsymbol{x}) = \begin{cases} (1-\alpha-\beta)\cdot\eta(\boldsymbol{x}) + \beta & \text{if (CCN)} \\ (1-2\tau)\cdot\eta(\boldsymbol{x}) + \tau & \text{if (UN)} \end{cases} \tag{2}$$

We consider loss functions that have the margin property: $\ell(y, f(x)) = \psi(yf(x))$, where $f : \mathbb{R}^d \to \mathbb{R}$ is a scorer, and $g(\boldsymbol{x}) = sign(f(\boldsymbol{x}))$ is the predictor. Let $f^* = \arg\min_{f\in\mathcal{F}} \mathbb{E}_{X,Y}\psi(Yf(X))$ and $\tilde{f}^* = \arg\min_{f\in\mathcal{F}} \mathbb{E}_{X,\tilde{Y}}\psi(\tilde{Y}f(X))$ denote the minimisers under the clean and noisy distributions, under model-class $\mathcal{F}$.

Definition 3. Uniform Noise robustness ([14])
Empirical risk minimization under loss function ℓ is said to be noise-tolerant if $\mathbb{P}_{X,Y}(g^*(X) = Y) = \mathbb{P}_{X,Y}(\tilde{g}^*(X) = Y)$.

Theorem 1. Sufficient conditions for robustness to uniform noise
Under uniform noise $\tau < 0.50$, and a margin loss function, $\ell(y, f(x)) = \psi(yf(x))$ satisfying: $\psi(f(x)) + \psi(-f(x)) = K$ for a positive constant K, we have that $\tilde{g}^*(x) = sign(\tilde{f}^*(x))$ obtained from: $\tilde{f}^* = \arg\min_{f\in\mathcal{F}} \mathbb{E}_{X,\tilde{Y}}\psi(\tilde{Y}f(X))$ is robust to uniform noise.

For the proof see Appendix 8.2. Several loss functions satisfy this, such as: the *square*, *unhinged* (linear), *logistic*, and more. We now introduce our definition of anchor points[1].

Definition 4. (Anchor Points) *An instance $\boldsymbol{x}$ is called an anchor point if we are provided with its true posterior $\eta(\boldsymbol{x})$. Let $\mathcal{A}_s^k$ denote a collection of k anchor points, with $\eta(\boldsymbol{x}) = s \;\forall \boldsymbol{x} \in \mathcal{A}_s^k$. Furthermore, let us also define $\mathcal{A}_{s,\delta}^k$, to imply that $\eta(\boldsymbol{x}_i) = s + \epsilon_i$, for $\epsilon_i \sim \mathbb{U}([-\delta,\ \delta])$, with $0 \le \delta \ll 1$ (respecting $0 \le \eta(\boldsymbol{x}) \le 1$). Also let $\mathcal{A}_{s,\delta} = \mathcal{A}_{s,\delta}^1$.*

$$\begin{array}{ccccc} \mathcal{A}_1^k & \to & \eta(\boldsymbol{x}) = 1 & \to & \tilde{\eta}(\boldsymbol{x}) = 1-\alpha \\ \mathcal{A}_{1/2}^k & \to & \eta(\boldsymbol{x}) = 1/2 & \to & \tilde{\eta}(\boldsymbol{x}) = \dfrac{1-\alpha+\beta}{2} \\ \mathcal{A}_0^k & \to & \eta(\boldsymbol{x}) = 0 & \to & \tilde{\eta}(\boldsymbol{x}) = \beta \end{array}$$

The cases we will be referring to are shown to the right. The first and last, $\mathcal{A}_1^k$ and $\mathcal{A}_0^k$, have been used in the past in different scenarios. In this work we will make use of the second case, $\mathcal{A}_{1/2}^k$.

[1] Different notions -related to our definition- of anchor points have been used before in the literature under different names. We review their uses and assumptions in Section 4

3 Hypothesis Tests based on anchor points

In this section we introduce our framework for devising hypothesis tests to examine the presence of class-conditional label noise in a given dataset (with uniform noise, as the alternative), assuming we are provided with an anchor point(s). Our procedure is based on a two-sided *z-test* (see for example Chapter 8 of [33]) with a simple null hypothesis, and a composite alternative hypothesis (Eq.5). We first define the distribution under the null hypothesis (Eq.6), and under the alternative hypothesis (Eq.7), when provided with one strict anchor point ($\eta(x) = {}^1\!/\!{}_2$). In this setting, for a fixed *level of significance* (Type I error) (Eq.8), we first derive a region for retaining the null hypothesis (Eq.9), and then we analyse the *power* (Prop.1) of the test (where we have that Type II Error = 1 - *power*). We then extend the approach to examine scenarios that include: (1) having multiple strict anchors ($\eta(x_i) = {}^1\!/\!{}_2,\ \forall i \in [k],\ k > 1$), (2) having multiple relaxed anchors ($\eta(x_i) \approx {}^1\!/\!{}_2,\ \forall i \in [k],\ k > 1$), and (3) having no anchors.

With the application of the *delta method* (See for example Chapter 3 of [15]) on Eq.1, we can get an asymptotic distribution for the predictive posterior:

$$\sqrt{n}(\hat{\eta}(\boldsymbol{x}) - \eta(\boldsymbol{x})) \xrightarrow{D} \mathcal{N}\left(0,\ \big(\eta(\boldsymbol{x})(1-\eta(\boldsymbol{x}))\big)^2 \cdot \boldsymbol{x}^\top \boldsymbol{I}_{\theta_0}^{-1} \boldsymbol{x}\right) \tag{3}$$

This would not work in the case of $\eta(\boldsymbol{x}) \in \{0,\ 1\}$, so instead we work with ${}^1\!/\!{}_2$. Which, together with the approximation of the Fisher-Information matrix with the empirical Hessian, we get:

$$\hat{\eta}(\boldsymbol{x}) \xrightarrow{D} \mathcal{N}\left(\frac{1}{2},\ \frac{1}{16} \cdot \boldsymbol{x}^\top \hat{H}_n \boldsymbol{x}\right) \tag{4}$$

where $\hat{H}_n = (X^\top D X)^{-1}$, where D is a diagonal matrix, with $D_{ii} = \hat{\eta}_i(1-\hat{\eta}_i)$, where $\hat{\eta}_i = \sigma(\boldsymbol{x}_i^\top \hat{\theta})$.

For the settings: $(\mathcal{D},\ \mathcal{A}^k_{1/2})$ and $(\mathcal{D}_\tau,\ \mathcal{A}^k_{1/2})$, for an $\boldsymbol{x} \in \mathcal{A}^k_{1/2}$ we get: $\tilde{\eta}(\boldsymbol{x}) = \frac{1}{2}$. While for $(\mathcal{D}_{\alpha,\beta},\ \mathcal{A}^k_{1/2})$ we get: $\tilde{\eta}(\boldsymbol{x}) = \frac{1-\alpha+\beta}{2}$. Note that under $(\mathcal{D}_\tau,\ \mathcal{A}^k_{1/2})$, we also have $\big(\tilde{\eta}(\boldsymbol{x})(1-\tilde{\eta}(\boldsymbol{x}))\big)^2 = \frac{1}{16}$ similarly to $(\mathcal{D},\ \mathcal{A}^k_{1/2})$.

3.1 A Hypothesis Test for Class-Conditional Label Noise

We now define our null hypothesis ($\mathcal{H}_0$) and (implicit) alternative hypothesis ($\mathcal{H}_1$) as follows:

$$\mathcal{H}_0 : \alpha = \beta \qquad \& \qquad \mathcal{H}_1 : \alpha \neq \beta \tag{5}$$

Under the null and the alternative hypotheses, we have the following distributions for the estimated posterior of the anchor:

$$\begin{aligned} \mathcal{H}_0 : \hat{\eta}(\boldsymbol{x}) &\sim \mathcal{N}\left(\frac{1}{2},\ \frac{1}{16} \cdot \boldsymbol{x}^\top \hat{H} \boldsymbol{x}\right) \\ &= \mathcal{N}\left(\frac{1}{2},\ v(\boldsymbol{x})\right) && (6) \\ \mathcal{H}_1 : \hat{\eta}(\boldsymbol{x}) &\sim \mathcal{N}\left(\frac{1+\alpha-\beta}{2},\ \tilde{v}(\boldsymbol{x})\right) && (7) \end{aligned}$$

where

$$\tilde{v}(\boldsymbol{x}) = \frac{\big((1-\alpha+\beta)(\beta-\alpha)\big)^2}{16} \cdot \boldsymbol{x}^\top \hat{\tilde{H}} \boldsymbol{x}$$

Level of Significance and Power of the test The *level of significance* (also known as Type I Error) is defined as follows:

$$a \ = \ \mathbb{P}(\text{reject } \mathcal{H}_0 \ \mid \ \mathcal{H}_0 \text{ is True}) \tag{8}$$

Rearranging Eq.6 we get: $\frac{\hat{\eta}(\boldsymbol{x}) - 1/2}{\sqrt{v(\boldsymbol{x})}} \sim \mathcal{N}(0,\ 1)$, under the null. Which for a chosen level of *significance* (a) allows us to define a region of retaining the null $\mathcal{H}_0$. We let $z_{a/2}$ and $z_{1-a/2}$ denote the lower and upper critical values for retaining the null at a level of significance of a.

Retain $\mathcal{H}_0$ if:

$$z_{a/2} \cdot \sqrt{v(x)} + 1/2 \ \leq \ \hat{\eta}(x) \ \leq \ z_{1-a/2} \cdot \sqrt{v(x)} + 1/2 \tag{9}$$

Using the region of retaining the null hypothesis, we can now derive the *power* of the test.

Proposition 1. Power of the test *(See Appendix 8.3 for the full derivation.) Under the distributions for the estimated posterior under the null and alternative hypotheses in Eqs.6&7, based on the definition of the hypotheses in Eq.5, the test has power:* $\mathbb{P}($*reject* $\mathcal{H}_0 \mid \mathcal{H}_0$ *is False*$) = 1 - b_1$*, where:*

$$b_1 = \Phi\left(\frac{z \cdot \sqrt{v(x)} + \frac{\beta-\alpha}{2}}{\sqrt{\tilde{v}(x)}}\right) - \Phi\left(\frac{-z \cdot \sqrt{v(x)} + \frac{\beta-\alpha}{2}}{\sqrt{\tilde{v}(x)}}\right) \tag{10}$$

3.2 Multiple Anchor Points

In this section we discuss how the properties of the test change in the setting where multiple anchors points are provided.

Let $\hat{\eta}_i$ correspond to the ith instance in $\mathcal{A}^k_{1/2}$. Then for $\bar{\eta} = \frac{1}{k}\sum_{i=1}^k \hat{\eta}_i$ we have:

$$\bar{\eta} \sim \mathcal{N}\left(\frac{1}{2},\ \frac{1}{16}\cdot \bar{\boldsymbol{x}}^\top H \bar{\boldsymbol{x}}\right)$$

where $\bar{\boldsymbol{x}} = \frac{1}{k}\sum_{i=1}^k \boldsymbol{x}_i$ with $\boldsymbol{x}_i \in \mathcal{A}^k_{1/2}\ \forall i$. For the full derivation see Appendix 8.4.

Anchors chosen at random We have that $\boldsymbol{x} \in \mathcal{A}^k_{1/2} \to \boldsymbol{x}^\top \beta_0 = 0$, so that for an orthonormal basis $\boldsymbol{U}$, $\boldsymbol{x} = \boldsymbol{U}\boldsymbol{r}$. Without loss of generalisation we let $\boldsymbol{U}_{:,0} = \frac{\beta_0}{\|\beta_0\|_2}$, and therefore $\eta(x) = 1/2 \to \boldsymbol{r}_0 = 0$. In words: $\forall \boldsymbol{x} \in \mathcal{A}^k_{1/2}$ we have that $\boldsymbol{x}$'s component in the direction of β_0 is 0.

Now we make the assumption that $\boldsymbol{x}$'s are random with $r_j \sim \mathbb{U}([-c,\ c])$. Therefore, $\mathbb{E}r_j = 0$, and $\mathbb{V}r_j = \frac{c^2}{3}$. In the following we use the subscript S in the operator $\mathbb{E}_S$ to denote the randomness in choosing the set $\mathcal{A}$. In words: we assume that the set $\mathcal{A}^k_{1/2}$ is chosen uniformly at random from the set of all anchor points.

Combining these we get:

$$\begin{aligned}\mathbb{E}_S v(x) &= \mathbb{E}_S x^\top H x = \mathbb{E}_S r^\top U H U^\top r \\ &= \frac{dc^2}{3}\cdot tr(UHU^\top) = \frac{dc^2 q}{3}\end{aligned}$$

where $q = tr(H)$. While for k anchor points chosen independently at random, we get:

$$\begin{aligned}\mathbb{E}_S v(\bar{x}) &= \mathbb{E}_S\left[\frac{1}{k^2}\sum_{i,j}^{k} x_i^\top H x_j\right] \\ &= \mathbb{E}_S\left[\frac{1}{k^2}\sum_{i,j}^{k} r_i^\top U H U^\top r_j\right] \\ &= \frac{dc^2}{3k}\cdot tr(UHU^\top) = \frac{dc^2 q}{3k}\end{aligned}$$

Following the same derivation as above we get:

$$b_k{=}\Phi\left(\frac{z\cdot\sqrt{v(\bar{x})}+\dfrac{\beta-\alpha}{2}}{\sqrt{\tilde{v}(\bar{x})}}\right) - \Phi\left(\frac{-z\cdot\sqrt{v(\bar{x})}+\dfrac{\beta-\alpha}{2}}{\sqrt{\tilde{v}(\bar{x})}}\right)$$

If we let $v = \mathbb{E}_S v(x)$ (similarly $\tilde{v} = \mathbb{E}_S \tilde{v}(x)$), then we have seen that $\mathbb{E}_S v(\bar{x}) = \frac{v}{k}$ (Reminder: expectations are with respect to the randomness in picking the anchor points). Then we have:

$$\frac{b_k}{b_1} = \frac{\Phi\left(\dfrac{z\sqrt{v}+h\sqrt{k}}{\sqrt{\tilde{v}}}\right) - \Phi\left(\dfrac{-z\sqrt{v}+h\sqrt{k}}{\sqrt{\tilde{v}}}\right)}{\Phi\left(\dfrac{z\sqrt{v}+h}{\sqrt{\tilde{v}}}\right) - \Phi\left(\dfrac{-z\sqrt{v}+h}{\sqrt{\tilde{v}}}\right)} \leq 1 \qquad (11)$$

with $h = \frac{\beta - \alpha}{2}$.

3.3 Multiple Relaxed Anchors-Points

In this section we see how the properties of the test change in the setting where the anchors do not have a perfect $\eta(\boldsymbol{x}) = 1/2$. We now consider the case of $\mathcal{A}^k_{1/2,\delta}$. Let $\boldsymbol{x}$ be such that $\eta(\boldsymbol{x}) = \frac{1}{2} + \epsilon$, where $\epsilon \sim \mathbb{U}([-\delta, \delta])$ with $0 < \delta \ll 1$. (Note: by definition $\delta \leq 1/2$.)

For one instance we have the following: $\mathbb{E}_{\hat{\theta}}\hat{\eta} = 1/2 + \epsilon$, and $\mathbb{E}_S\mathbb{E}_{\hat{\theta}}\hat{\eta} = 1/2$

For the variance component we have: $(\hat{\eta}(1-\hat{\eta}))^2 = \left(\left(\frac{1}{2}+\epsilon\right)\left(\frac{1}{2}-\epsilon\right)\right)^2 \approx \frac{1}{16} - \frac{\epsilon^2}{2}$, ignoring terms of order higher than ϵ^2, under the assumption that $\delta \ll 1$.

Under the *law of total variance* we have:

$$\begin{aligned}
\mathbb{V}(\eta) &= \mathbb{E}\left(\mathbb{V}\left(\eta \mid \epsilon\right)\right) + \mathbb{V}\left(\mathbb{E}\left(\eta \mid \epsilon\right)\right) \\
&= \mathbb{E}\left(\left(\frac{1}{16} - \frac{\epsilon^2}{2}\right) \cdot \boldsymbol{x}^\top H \boldsymbol{x}\right) + \mathbb{V}\left(\frac{1}{k}\sum_{i=1}^{k}\hat{\eta}_i\right) \\
&= \left(\frac{1}{16} - \frac{\delta^2}{6}\right) \cdot \boldsymbol{x}^\top H \boldsymbol{x} + \mathbb{V}\left(\frac{1}{2} + \frac{1}{k}\sum_{i=1}^{k}\epsilon_i\right) \\
&= \left(\frac{1}{16} - \frac{\delta^2}{6}\right) \cdot \boldsymbol{x}^\top H \boldsymbol{x} + \frac{\delta^2}{3k} \qquad (12)
\end{aligned}$$

For the full derivation see Appendix 8.6. Finally, bringing everything together and ignoring δ^2 terms we get:

$$\begin{aligned}
\bar{\eta} &\sim \mathcal{N}\left(\frac{1}{2}, \left(\frac{1}{16} - \frac{\delta^2}{6}\right) \cdot \bar{\boldsymbol{x}}^\top H \bar{\boldsymbol{x}}\right) \\
&\approx \mathcal{N}\left(\frac{1}{2}, \frac{1}{16} \cdot \bar{\boldsymbol{x}}^\top H \bar{\boldsymbol{x}}\right)
\end{aligned}$$

3.4 What if we have no anchor points?

We have shown that we can relax the hard constraint on the anchor points to be exactly $\eta = 1/2$, to $\eta \approx 1/2$. It is natural then to ask if we need anchor points at all. If instead we were to sample points at random, then we would have the following: $\mathbb{E}_{p(X)}\eta(X) = \pi$. The importance of needing for set of anchor points, either $\mathcal{A}^k_{1/2}$ or $\mathcal{A}^k_{1/2,\delta}$, is that, the anchor points would be centered around a known value $1/2$, as opposed to having no anchor points and sampling at random, where the anchor points would end up being centered around π. Knowledge of the class priors could allow for a different type of hypothesis tests to asses the presence of label noise. We do not continue this discussion in the main document as it relies on different type of information, but provide pointers in the Appendix 7.8.

3.5 Practical Considerations & Limitations

Beyond Logistic Regression Our approach relies on the asymptotic properties of MLE estimators, and specifically of Logistic Regression. More complex models can be constructed in a similar fashion through polynomial feature expansion. However the extension of these tests to richer model-classes, such as Gaussian Processes, remains open.

Multi-class classification Multi-class classification setting can be reduced to *one-vs-all*, *all-vs-all*, or more general error-correcting output codes setups as described in [23], which rely on multiple runs of binary classification. In these settings then we could apply the proposed framework. The challenge would then be how to interpret $\eta = {}^1/_2$.

Finding anchor points While it might not be straightforward for the user to provide instances whose true posterior is $\eta(\boldsymbol{x}) = {}^1/_2$, we do show how this can be relaxed, by allowing $\eta(\boldsymbol{x}) \approx {}^1/_2$. We then show how multiple anchor points can be stacked, improving the properties of the test.

Model Misspecification Our work relies on properties of the MLE and its asymptotic distribution (Eq. 1). These assume the model is *exactly* correct. Similarly, under the null in the scenario of $\alpha = \beta > 0$, we are at risk of model misspecification. This is not a new problem for Maximum Likelihood estimators, and one remedy is the so-called *Huber Sandwich Estimator* [34] which replaces the Fisher Information Matrix, with a more robust alternative.

Instance-dependent Noise (IDN) In IDN the probability of label flipping depends on the features. It can be seen as a generalisation over UN (which is unbiased under mild conditions (See Theorem 1) and CCN (where learning is in general biased). Our theoretical framework for CCN serves as a starting point to devise tests of IDN.

4 Related Work

Previous works have focused on the importance of (automatic) data preparation and data quality assessment [24, 36–38]. These data quality measures refer to aspects such as the presence of noise in data, missing values, outliers, imbalanced classes, inconsistency, redundancy, timeliness and more [36, 38]. Within this context, in this work we focus on label noise and, in particular, assessing the presence of class-conditional label noise, as opposed to uniform label noise. Related approaches include the identification specific corrupted instances, or distilled examples, and the direct estimation of the noise rates. These are discussed below.

Noisy examples As presented in [11,26], the aim is to identify the *specific* examples that have been inflicted with noise. This is a non-trivial task unless certain assumptions can be made about the per-class distributions, and their shape. For example, if we can assume that the supports of the two classes do not overlap (i.e. $\eta(x)(1-\eta(x)) \in \{0,\ 1\}\ \forall x$), then we can identify mislabelled instances using per-class densities. If this is not the case, then it would be difficult to differentiate between a mislabelled instance and an instance for which $\eta(x)(1-\eta(x)) \in (0,1)$. A different assumption could be uni-modality, which would again provide a prescription for identifying mislabelled instances through density estimation tools.

Distilled examples The authors in [16] go in the opposite direction by trying to identify instances that *have not been corrupted* → the *distilled examples.* As a first step the authors assume knowledge of an upper-bound[2] (Theorem 2 of [16]) which allows them to define sufficient conditions for identifying whether an instance is *clean.* As a second step they aim at estimating the (local) noise rate based on the neighbourhood of an instance (Theorem 3 of [16]).

Anchor points and perfect samples Finally, we can aim to directly estimate noise rates (or general distortions) while training [17,39]. A common approach is to proceed by correcting the loss to be minimised, by introducing the notion of a *mixing matrix* $\boldsymbol{M} \in [0,1]^{c\times c}$, where $M_{i,j} = \mathbb{P}(\tilde{y} = \boldsymbol{e}^j \mid y = \boldsymbol{e}^i)$ [8]. Using these formulations, we are in a position where, if we have access to $\boldsymbol{M}$, we can correct the training procedure to obtain an unbiased estimator. However, $\boldsymbol{M}$ is rarely known and difficult to estimate. Works on estimating $\boldsymbol{M}$ rely on having access to *perfect samples* and can be traced back to [3], and the idea was later adapted and generalised in [4,5,17] to the multi-class setting. Interestingly, in [1] authors do not explicitly define these perfect samples, but rather assume they do exist in a large enough (validation) dataset $\boldsymbol{X}'$ – obtaining good experimental results. Similarly, [18] also work by not explicitly requiring anchor points, but rather assuming their existence.

5 Experiments

In order to illustrate the properties of the tests, for the experiments we consider a synthetic dataset where the per-class distributions are Gaussians, with means $[1,\ 1]^\top$ and $[-1,\ -1]^\top$, with identity as scale. For this setup we know that anchor points should lie on the line $y = -x$, and draw them uniformly at random $x \in [-4,\ 4]$. We analyse the following parameters of interest:

1. $N \in [500,\ 1000,\ 2000,\ 5000]$: the training sample size.
2. $(\alpha - \beta) \in [-0.05,\ 0.10,\ 0.20]$: the difference between the per-class noise rates.
3. $k \in [1,\ 2,\ 4,\ 8,\ 16,\ 32]$: the number of anchor points.

[2] The paper aims at tackling instance-dependent noise.

4. $\delta \in [0,\ 0.05,\ 0.10]$: how relaxed the anchor points are: $\eta(x) \in [0.50-\delta,\ 0.50+\delta]$.

For all combinations of N and $(\alpha - \beta)$ we perform 500 runs. In each run, we generate a clean version of the data $\mathcal{D}$, and then proceed by corrupting it to obtain a separate version: $\mathcal{D}_{\alpha,\beta}$. For both datasets, we fit a Logistic Regression model. We sample both the anchor points and relaxed anchor points. Finally, we then compute the z-scores, and subsequently the corresponding p-values[3].

The box-plots should be read as follows: $Q1$, $Q2$ & $Q3$ separate the data into 4 equal parts. The inner box starts (at the bottom) at $Q1$ and ends (at the top) at $Q3$, with the horizontal line inside denoting the median ($Q2$). The whiskers extend to show $Q1 - 1.5 \cdot IQR$, and $Q3 + 1.5 \cdot IQR$. IQR denotes the *Interquartile Range* and $IQR = Q3 - Q1$.

In Figures 1, 2 and 3 we have the following: moving to the right we increase the relaxation of anchor points, and moving downwards we increase the training sample-size. On the subplot level, on the x-axis we vary the number of anchor points, and on the y-axis we have the p-values. In all subplots we indicate with a red dashed line the mark of 0.10, and with a blue one the mark of 0.05, which would serve as rejection thresholds for the null hypothesis.

The experiments are illustrative of the claims made earlier in the paper. Below we discuss the findings in the experiments and what they mean with regards to Type I and Type II errors. We discuss these points in two parts; we first discuss the effect on sample size (N), difference in noise rates ($|\alpha - \beta|$) and number of anchor points (k).

Size of training set (N) As the size of training set (N) increases, the power increases. This can be seen Figures 1, 2 & 3. By moving down the first column, and fixing a value for k, where N increases, we see the range of the purple box-plots decreasing, and essentially a larger volume of tests falling under the cut-off levels of significance (red and blue dashed lines). This is expected given that the variance of the MLE $\hat{\theta}_{MLE}$ vanishes as N increases, as is seen in Eq.1 and discussion underneath it.

Difference in noise rates ($|\alpha - \beta|$) As $|\alpha - \beta|$ increases, the power increases. This can be seen in Figures 1, 2 & 3, by fixing a particular subplot in the first column (for example, top-left one), and a value for k, we see again that the volume moves down. As presented in Eq.10, as $\beta - \alpha$ increases, the power also increases.

[3] What we have so far presented is aligned with the Neyman-Pearson theory of hypothesis testing. We have shown how to utilise anchor points to obtain the p-value – a continuous measure of evidence against the null hypothesis- and then leverage the implicit *alternative hypothesis* of class-conditional noise and a *significance level* to analyse the *power* of the test. In this case, the p-value is the basis of formal decision-making process of rejecting, or failing to reject, the null hypothesis. Differently, in Fisher's theory of significance testing, the p-value is the end-product [35]. Both the p-value and the output of the test can be used as part of a broader decision process that considers other important factors.

Number of anchor points (k) The same applies to the number of anchor points – as the number of anchor points (k) increases, the power of the test increases. This can be seen in all three figures by focusing in any subplot in the first column, and considering the purple box-plots moving to the right. In Eq.11 we see effect of k on the power.

In all three discussions above we focused on the first column of each of the figures – which shows results from experiments on strict anchors. What we also observe in this case (the first column of all figures) is that the p-values follow the uniform distribution under the null (as expected, given the null hypothesis is true) – shown by the green box-plots. Therefore the portion of Type I Errors $= a$ (the level of significance Eq.8). When we relax the requirements for strict anchors to allow for values close to $^1/_2$, we introduce a bias in the lower and upper bounds in Eq.9 of $+\epsilon$. While $\mathbb{E}\epsilon = 0$ this shift on the boundaries of the retention region will increase Type I Error. On the other hand, in Eq.12 we see how this bias decreases as you increase the number of anchor points. Both of these phenomena are also shown experimentally by looking at the latter two columns of the figures.

Anchor point relaxation (δ) Lastly, we examine the effect of relaxing the strictness of the anchors (δ), $\eta(x) \in [0.50 - \delta,\ 0.50 + \delta]$ on the properties of the test. As just discussed we see that as we increase the number of anchor points Type I Error decreases (volume of green box-plots under each of the cut-off points). We also observe that, as compared to only allowing strict anchors, the power is not affected significantly – with the effect decreasing as the number of anchor points increases. Furthermore, in the latter two columns we also observe the phenomena mentioned in the discussion concerning the first column only.

6 Conclusion & Future Work

In this work we introduce the first statistical hypothesis test for class-conditional label noise. Our approach requires the specification of anchor points, i.e. instances whose labels are highly uncertain under the true posterior probability distribution, and we show that the test's significance and power is preserved over several relaxations on the requirements for these anchor points. Our experimental analysis, which confirms the soundness of our test, explores many configurations of practical interest for practitioners using this test. Of particular importance for practitioners, since anchor specification is under their control, is the high correspondence shown theoretically and experimentally between the number of anchors and test significance.

Future work will cover both theoretical and experimental components. On the theoretical front, we are interested in understanding the test's value under a richer set of classification models, and further relaxing requirements on true posterior uncertainty for anchor points. Experimentally, we are particularly interested in applying the tests to diagnostically challenging healthcare problems and utilising clinical experts for anchor specification.

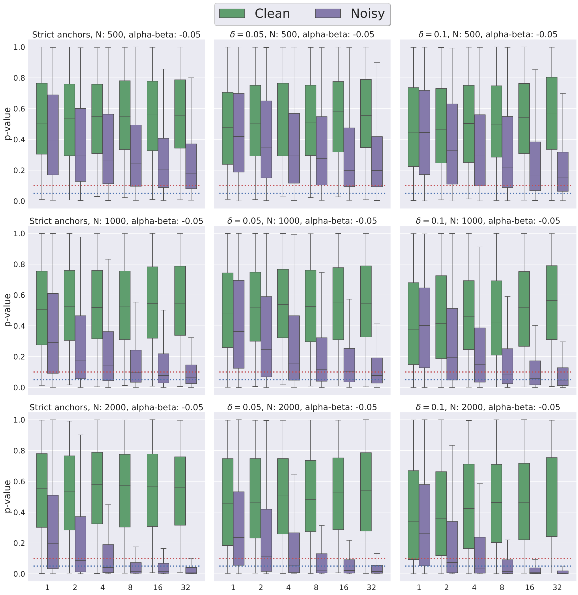

Fig. 1: Fixed $|\beta - \alpha| = 0.05$. Red dotted line at 0.10, and blue at 0.05.

Acknowledgements

This work was funded by the UKRI Turing AI Fellowship EP/V024817/1.

References

1. Patrini, Giorgio, et al. "Making deep neural networks robust to label noise: A loss correction approach." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.
2. Wang, Daixin, Peng Cui, and Wenwu Zhu. "Structural deep network embedding." Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining. 2016.
3. Blanchard, Gilles, et al. "Classification with asymmetric label noise: Consistency and maximal denoising." Electronic Journal of Statistics 10.2 (2016): 2780-2824.
4. Menon, Aditya, et al. "Learning from corrupted binary labels via class-probability estimation." International conference on machine learning. PMLR, 2015.
5. Liu, Tongliang, and Dacheng Tao. "Classification with noisy labels by importance reweighting." IEEE Transactions on pattern analysis and machine intelligence 38.3 (2015): 447-461.
6. Patrini, Giorgio. "Weakly supervised learning via statistical sufficiency." (2016).
7. Van Rooyen, Brendan. "Machine learning via transitions." (2015).
8. Cid-Sueiro, Jesús. "Proper losses for learning from partial labels." Advances in neural information processing systems 25 (2012).

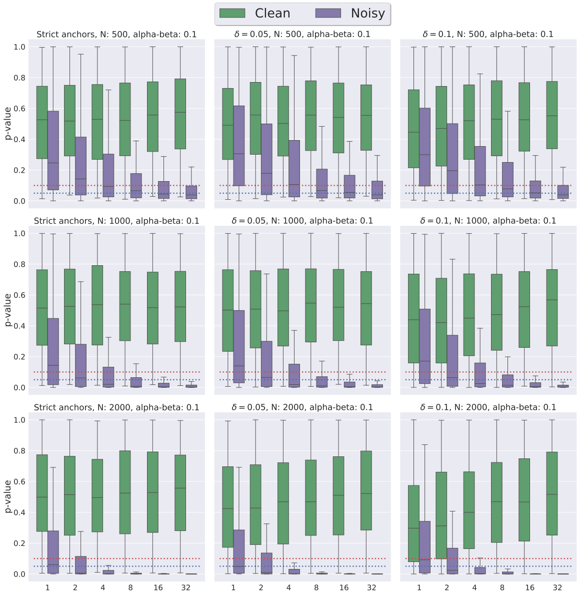

Fig. 2: Fixed $|\beta - \alpha| = 0.10$. Red dotted line at 0.10, and blue at 0.05.

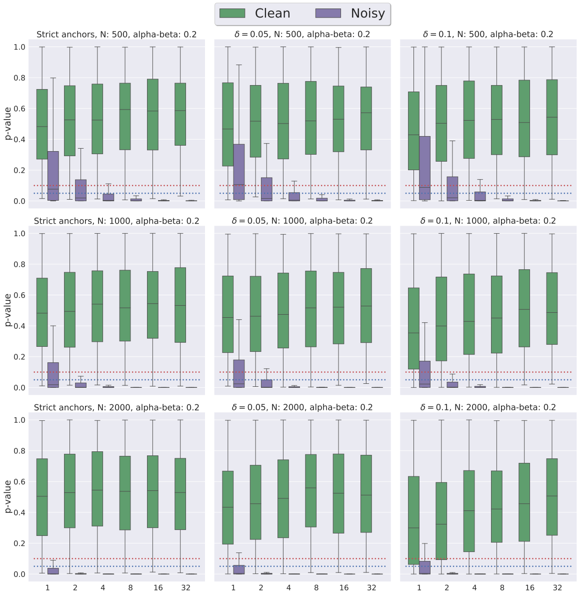

Fig. 3: Fixed $|\beta - \alpha| = 0.20$. Red dotted line at 0.10, and blue at 0.05.

9. Cid-Sueiro, Jesús, Darío García-García, and Raúl Santos-Rodríguez. "Consistency of losses for learning from weak labels." Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, Berlin, Heidelberg, 2014.
10. Perelló-Nieto, Miquel, Raúl Santos-Rodríguez, and Jesús Cid-Sueiro. "Adapting supervised classification algorithms to arbitrary weak label scenarios." International Symposium on Intelligent Data Analysis. Springer, Cham, 2017.
11. Northcutt, Curtis, Lu Jiang, and Isaac Chuang. "Confident learning: Estimating uncertainty in dataset labels." Journal of Artificial Intelligence Research 70 (2021): 1373-1411.
12. Frénay, Benoît, and Michel Verleysen. "Classification in the presence of label noise: a survey." IEEE transactions on neural networks and learning systems 25.5 (2013): 845-869.
13. Poyiadzi, Rafael, et al. "The Weak Supervision Landscape." 2022 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops). IEEE, 2022.
14. Ghosh, Aritra, Naresh Manwani, and P. S. Sastry. "Making risk minimization tolerant to label noise." Neurocomputing 160 (2015): 93-107.
15. Van der Vaart, Aad W. Asymptotic statistics. Vol. 3. Cambridge university press, 2000.
16. Cheng, Jiacheng, et al. "Learning with bounded instance and label-dependent label noise." International Conference on Machine Learning. PMLR, 2020.
17. Perello-Nieto, Miquel, et al. "Recycling weak labels for multiclass classification." Neurocomputing 400 (2020): 206-215.
18. Xia, Xiaobo, et al. "Are anchor points really indispensable in label-noise learning?." Advances in Neural Information Processing Systems 32 (2019).
19. Bedrick, Edward J., Ronald Christensen, and Wesley Johnson. "A new perspective on priors for generalized linear models." Journal of the American Statistical Association 91.436 (1996): 1450-1460.
20. Greenland, Sander. "Putting background information about relative risks into conjugate prior distributions." Biometrics 57.3 (2001): 663-670.
21. Gelman, Andrew, et al. "A weakly informative default prior distribution for logistic and other regression models." The annals of applied statistics 2.4 (2008): 1360-1383.
22. Garthwaite, Paul H., Joseph B. Kadane, and Anthony O'Hagan. "Statistical methods for eliciting probability distributions." Journal of the American statistical Association 100.470 (2005): 680-701.
23. Dietterich, Thomas G., and Ghulum Bakiri. "Solving multiclass learning problems via error-correcting output codes." Journal of artificial intelligence research 2 (1994): 263-286.
24. Lawrence, Neil D. "Data readiness levels." arXiv preprint arXiv:1705.02245 (2017).
25. Jindal, Ishan, Matthew Nokleby, and Xuewen Chen. "Learning deep networks from noisy labels with dropout regularization." 2016 IEEE 16th International Conference on Data Mining (ICDM). IEEE, 2016.
26. Northcutt, Curtis G., Tailin Wu, and Isaac L. Chuang. "Learning with confident examples: Rank pruning for robust classification with noisy labels." arXiv preprint arXiv:1705.01936 (2017).
27. Gebru, Timnit, et al. "Datasheets for datasets." Communications of the ACM 64.12 (2021): 86-92.
28. Sokol, Kacper, Raul Santos-Rodriguez, and Peter Flach. "FAT forensics: a python toolbox for algorithmic fairness, accountability and transparency." arXiv preprint arXiv:1909.05167 (2019).

29. Zhao, Liyue, Gita Sukthankar, and Rahul Sukthankar. "Incremental relabeling for active learning with noisy crowdsourced annotations." 2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing. IEEE, 2011.
30. Bacaicoa-Barber, Daniel, et al. "On the selection of loss functions under known weak label models." International Conference on Artificial Neural Networks. Springer, Cham, 2021.
31. Fergus, Robert, et al. "Learning object categories from google's image search." Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1. Vol. 2. IEEE, 2005.
32. Schroff, Florian, Antonio Criminisi, and Andrew Zisserman. "Harvesting image databases from the web." IEEE transactions on pattern analysis and machine intelligence 33.4 (2010): 754-766.
33. Casella, George, and Roger L. Berger. Statistical inference. Cengage Learning, 2021.
34. Freedman, David A. "On the so-called “Huber sandwich estimator” and “robust standard errors”." The American Statistician 60.4 (2006): 299-302.
35. Perezgonzalez, Jose D. "Fisher, Neyman-Pearson or NHST? A tutorial for teaching data testing." Frontiers in psychology 6 (2015): 223.
36. Gupta, Nitin, et al. "Data Quality Toolkit: Automatic assessment of data quality and remediation for machine learning datasets." arXiv preprint arXiv:2108.05935 (2021).
37. Afzal, Shazia, et al. "Data readiness report." 2021 IEEE International Conference on Smart Data Services (SMDS). IEEE, 2021.
38. Corrales, David Camilo, Agapito Ledezma, and Juan Carlos Corrales. "From theory to practice: A data quality framework for classification tasks." Symmetry 10.7 (2018): 248.
39. Chu, Zhendong, Jing Ma, and Hongning Wang. "Learning from Crowds by Modeling Common Confusions." AAAI. 2021.