Robust Learning of Multi-Label Classifiers under Label Noise

by

Himanshu Kumar, Naresh Manwani, Sastry P S

in

ACM India Joint International Conference on Data Science Management of Data
(*CODS-COMAD-2020*)

Report No: IIIT/TR/2020/-1

Centre for Visual Information Technology
International Institute of Information Technology
Hyderabad - 500 032, INDIA
January 2020

Robust Learning of Multi-Label Classifiers under Label Noise

Himanshu Kumar
EE, IISc Bangalore
Karnataka, India
khimanshu@iisc.ac.in

Naresh Manwani
IIIT Hyderabad
Telangana, India
naresh.manwani@iiit.ac.in

P. S. Sastry
EE, IISc Bangalore
Karnataka, India
sastry@iisc.ac.in

ABSTRACT

In this paper, we address the problem of robust learning of multi-label classifiers when the training data has label noise. We consider learning algorithms in the risk-minimization framework. We define what we call symmetric label noise in multi-label settings which is a useful noise model for many random errors in the labeling of data. We prove that risk minimization is robust to symmetric label noise if the loss function satisfies some conditions. We show that Hamming loss and couple of surrogates of Hamming loss satisfy these sufficient conditions and hence are robust. By learning feed-forward neural networks on some benchmark multi-label datasets, we provide empirical evidence to illustrate our theoretical results on robust learning of multi-label classifiers under label noise.

KEYWORDS

multi-label, neural networks, label noise, robust losses

ACM Reference Format:
Himanshu Kumar, Naresh Manwani, and P. S. Sastry. 2020. Robust Learning of Multi-Label Classifiers under Label Noise. In *7th ACM IKDD CoDS and 25th COMAD (CoDS COMAD 2020), January 5–7, 2020, Hyderabad, India.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3371158.3371169

1 INTRODUCTION

Traditionally classification algorithms are formulated to predict only one class label for each input pattern. The underlying assumption is that object categories are disjoint. However, there are many classification scenarios where such an assumption is restrictive. For example, in image annotation, each image may simultaneously belong to multiple classes [4]. In document classification, each document can belong to multiple topics [31]. In gene classification, each gene may belong to multiple function classes [35]. Such problems where the classification algorithm is required to predict a set of labels (rather than a unique label) are termed as multi-label problems. Multi-label classifier learning is an interesting problem in current research [36].

Many of the algorithms learn a classifier assuming that class labels in the training data are correct, which may not always be true. Presence of incorrect labels in training data is referred to as label noise. Robust learning of classifiers refers to the problem of learning a classifier that generalizes well under the underlying (and unknown) noise-free distribution given training data with noisy labels. This is an important and challenging problem [8]. Many times labels of training data are corrupted due to insufficient information available to expert labelers, unavoidable human errors, subjective biases, or inherent noise in the data generation process. Label noise in training data is more prevalent in recent times as many datasets are prepared through crowd-sourcing. Such label noise is particularly unavoidable in multi-label problems due to the inherent subjective biases in deciding on which all labels are appropriate for a given object.

In this paper, we study multi-label classification under label noise. We define a uniform/symmetric label noise model in multi-label settings. Symmetric label noise model described here is reasonable and can capture many subjective biases or random human errors while labeling. We then show that classifier learning through risk minimization is robust to symmetric label noise if the loss function satisfies what we call a symmetry condition. We show that the standard hamming loss and a classification calibrated surrogate of hamming loss fulfill this condition. We also present experimental results to illustrate the robust learning of classifiers under such symmetric losses.

2 RELATED WORK

There have been many methods proposed for robust learning of classifiers in the presence of label noise. (Some recent surveys are [8, 25]). However, almost all the methods are applicable only for binary or multi-class classification problems. There are hardly any robustness results for learning multi-label classifiers.

Some data processing methods rely on detecting and removing (or correcting) noisy samples from training data, and many different heuristics are proposed for this [1, 5, 37]. There are also some methods which heuristically modify the existing algorithms to make them robust to label noise [2, 14, 15]. In addition to training data with noisy labels, we have some data with clean labels then one can use techniques from semi-supervised learning. For example, such a method is recently shown to be useful for learning with noisy labels in a multi-label problem [12]. Another general approach that is followed is to view the unknown 'true' class labels of training examples as latent or hidden variables. Assuming probabilistic noise models, one can estimate a generative or discriminative model for the classification problem using an EM-type algorithm [3, 17, 20, 33]. Similar techniques are proposed for making deep neural networks robust to label noise [24, 28]. In a multi-label setting, probabilistic models for the corruption of labels have been used in conjunction with a topic model for document classification [22]. While many of these methods are reported to perform well under label noise, there are no theoretical guarantees of robustness.

Recently, risk minimization techniques have been developed that are effective in tackling label noise in training data [19]. These

CoDS COMAD 2020, January 5–7, 2020, Hyderabad, India

ACM ISBN 978-1-4503-7738-6/20/01…$15.00
https://doi.org/10.1145/3371158.3371169

methods also provide theoretical guarantees about the robustness of the learned classifier. Robustness of risk minimization depends on the loss function used. It is shown, for binary classification case, that risk minimization under 0–1 loss is robust to symmetric or uniform label noise while that under any of the standard convex losses is not [18, 19]. Interestingly, unhinged loss, a convex loss that is not a convex potential, is robust to symmetric label noise [32]. Some of the robust risk minimization methods work as follows. Given a loss function (satisfying some properties) they perform risk minimization on a modified loss function which is provably robust. For constructing the modified loss function, one needs noise probabilities which are (implicitly or explicitly) estimated from the training data [21, 23, 26]. The theoretical guarantees of the robustness of these methods assume knowledge of the exact noise probabilities. Another approach is to find some sufficient conditions on the loss function so that risk minimization would be provably robust. Such results are proved for the binary classification case in [11] and for the multi-class case in [10]. In practice, many such risk minimization approaches are seen to be quite useful in tackling label noise in training data.

In the case of multi-label learning, robustness has multiple connotations. As each example is labeled with a set of class labels, certain combinations of class labels may rarely occur in the training set. Thus, we may not learn the proper correlations among the multiple labels. We may want classifier learning techniques that are robust in the sense of taking care of such rarely occurring label combinations. Recently some interesting approaches are proposed to tackle such issues [6, 34]. However, to the best of our knowledge, there are no theoretical results on the robustness to label noise in the multi-label setting.

The rest of the paper is organized as follows. In Section 3, we describe the problem setting of multi-label classification and risk minimization and explain symmetric label noise model for multi-label learning. In Section 4, we present our main theoretical result regarding the robustness of risk minimization and show that hamming loss is robust. In Section 5 we empirically demonstrate our theoretical results on robust learning under label noise. We present some discussion and conclusions in Section 6.

3 PROBLEM DESCRIPTION

Let $\mathcal{X} \subset \mathbb{R}^d$ be the instance space or feature vector space and let $\mathcal{Y} = [k] = \{1, \cdots k\}$ be the set of class labels. In multi-label classification, labels associated with an instance are a subset of $\mathcal{Y}$. This label-set associated with each instance, $\mathbf{x}$, can be represented by a vector $\mathbf{y_x} \in \{0,1\}^k$ where $(\mathbf{y_x})_q$, the q^{th} component of $\mathbf{y_x}$, is 1 if and only if q is one of the labels associated with this instance.

In a multi-label classifier learning problem, we are given training data, $S = \{(\mathbf{x}_1, \mathbf{y}_{\mathbf{x}_1}), \ldots, (\mathbf{x}_N, \mathbf{y}_{\mathbf{x}_N})\} \in (\mathcal{X} \times \{0,1\}^k)^N$, drawn *iid* according to an unknown distribution, $\mathcal{D}$, over $\mathcal{X} \times \{0,1\}^k$. We need to learn a classifier, $h : \mathcal{X} \to \{0,1\}^k$. We often represent a classifier as $\mathbf{h}(\mathbf{x}) = \text{pred} \circ \mathbf{f}(\mathbf{x})$ where $\mathbf{f} : \mathcal{X} \to C,\ C \subseteq \mathbb{R}^k$. (The classifier $\mathbf{h}$ predicts the class label given $\mathbf{f}(\mathbf{x}) \in \mathbb{R}^k$). Here the function 'pred' might be a simple rule such as thresholding each component of $\mathbf{f}(\mathbf{x})$. Even though the final classification decision on a feature vector $\mathbf{x}$ is $\text{pred} \circ \mathbf{f}(\mathbf{x})$, we use the notation of calling $\mathbf{f}$ itself as the classifier.

Risk minimization is a popular strategy for learning a classifier. Risk of a classifier $\mathbf{f}$ under loss function L is defined as

$$R_L(\mathbf{f}) = \mathbb{E}_{\mathcal{D}}[L(\mathbf{f}(\mathbf{x}), \mathbf{y_x})] = \mathbb{E}_{\mathbf{x},\mathbf{y_x}}[L(\mathbf{f}(\mathbf{x}), \mathbf{y_x})] \tag{1}$$

where $\mathbb{E}_{\mathcal{D}}$ denotes expectation with respect to the underlying distribution $\mathcal{D}$. The risk depends on the loss function and hence, often the above is called the L-risk of $\mathbf{f}(\mathbf{x})$. We want to learn a classifier that has minimum L-risk, namely,

$$\mathbf{f}^* = \arg\min_{\mathbf{f} \in \mathcal{F}} R_L(\mathbf{f}) \tag{2}$$

where $\mathcal{F}$ is a function class over which we are minimizing. The classifier we learn through risk minimization depends on our choice of a loss function.

When there is label noise, the learner does not have access to the clean training data (represented by S above) which is drawn according to distribution $\mathcal{D}$. Instead, we have access only to noisy data which is drawn according to a distribution $\mathcal{D}_\eta$. Distribution $\mathcal{D}_\eta$ is result of the underlying corruption process. Here we consider what we call symmetric or uniform noise which we define now.

The noisy training data available to the learner would be denoted by $S_\eta = \{(\mathbf{x}_n, \tilde{\mathbf{y}}_{\mathbf{x}_n}), n = 1, \cdots, N\}$ drawn according to distribution $\mathcal{D}_\eta$. Here $\tilde{\mathbf{y}}_{\mathbf{x}_n}$ is the noisy label of $\mathbf{x}_n$. A noise model specifies how the noisy labels relate to the true labels.

For multi label learning, we say noise is *symmetric* when:

$$(\tilde{\mathbf{y}}_{\mathbf{x}_n})_q = \begin{cases} (\mathbf{y}_{\mathbf{x}_n})_q & \text{with probability } (1-\eta_q) \\ 1 - (\mathbf{y}_{\mathbf{x}_n})_q & \text{with probability } \eta_q \end{cases} \tag{3}$$

where $\eta_q \in [0,1]$ is a constant, called the noise probability or noise rate for the q^{th} label. Under this noise model, each component of the binary vector $\mathbf{y_x}$ is independently corrupted with the corruption probability for the q^{th} component being η_q, $q = 1, \cdots, k$. We call this symmetric because, under noise corruption, the binary value, $(\mathbf{y_x})_q$, turning from 0 to 1 or 1 to 0 has the same probability. When $\eta_q = \eta\ \forall q$ then we call it a *uniformly symmetric label noise*. These are simple noise models but they are reasonable for modelling many random labelling errors in generating the training data. Since the symmetric label noise allows for different noise rates for different labels, it can handle cases where some categories are prone for higher level of confusion.

In presence of label noise, we observe samples only from the noisy distribution D_η and hence can only minimize L-risk with respect to $\mathcal{D}_\eta$. Let the L-risk of $\mathbf{f}$ under the noisy distribution be denoted by $R_L^\eta(\mathbf{f})$, which is defined by

$$R_L^\eta(\mathbf{f}) = \mathbb{E}_{\mathcal{D}_\eta}[L(\mathbf{f}(\mathbf{x}), \tilde{\mathbf{y}}_\mathbf{x})] = \mathbb{E}_{\mathbf{x},\tilde{\mathbf{y}}_\mathbf{x}}[L(\mathbf{f}(\mathbf{x}), \tilde{\mathbf{y}}_\mathbf{x})] \tag{4}$$

Let

$$\mathbf{f}_\eta^* = \arg\min_{\mathbf{f} \in \mathcal{F}} R_L^\eta(\mathbf{f}) \tag{5}$$

Given the noisy samples, S_η, we can only learn $\mathbf{f}_\eta^*$ through risk minimization.

Thus, the problem is to learn a classifier which minimizes risk under distribution $\mathcal{D}_\eta$ to have same performance to a classifier which minimizes risk over distribution $\mathcal{D}$. We next discuss it in context of noise-tolerance (robustness) property of loss functions.

Definition 1: Risk minimization under loss function L, is said to be *noise-tolerant* [19] if

$$\text{Prob}_{\mathcal{D}}[\text{pred} \circ \mathbf{f}^*(\mathbf{x}) = \mathbf{y_x}] = \text{Prob}_{\mathcal{D}}[\text{pred} \circ \mathbf{f}^*_\eta(\mathbf{x}) = \mathbf{y_x}]$$

When the above holds we also say the loss function is noise-tolerant.

Noise-tolerance is a very useful property. A loss function is noise-tolerant if the minimizers of risk under the noise-free and noisy distributions, both have the same probability of misclassification under the noise-free distribution. If we take the noisy labels as true labels and minimize risk, then if we are learning a classifier through risk minimization using a noise-tolerant loss function, then we can automatically take care of label noise. With a noise-tolerant loss function, we need not even know whether or not there is label noise.

4 ROBUSTNESS TO SYMMETRIC LABEL NOISE

In this section we prove a general sufficient condition for a loss function to be noise-tolerant for symmetric and uniformly symmetric label noise in the multi-label setting. This could be viewed as a generalization, to the multi-label setting, of similar results for single-label case [10, 11].

THEOREM 1. *Multi-label classifier learning under risk minimization is noise-tolerant to symmetric label noise if $\eta_q < \frac{1}{2}\ \forall q$ and the loss function, $L(\mathbf{f}(\mathbf{x}), \mathbf{y_x})$, satisfies the following conditions:*

C1. $L(\mathbf{f}(\mathbf{x}), \mathbf{y_x}) = \sum_{q=1}^{k} l((\mathbf{f}(\mathbf{x}))_q, (\mathbf{y_x})_q) \quad \forall \mathbf{x}, \mathbf{y_x}, \mathbf{f}$

C2. $l(f(\mathbf{x}), 0) + l(f(\mathbf{x}), 1) = C \quad \forall \mathbf{x}, f$

Here C is a constant greater than zero and $l(f(\mathbf{x}), (\mathbf{y_x})_q)$ is a loss function for binary classification. (Note that, here we have $\mathbf{f}(\mathbf{x}) \in \mathbb{R}^k$, $f(\mathbf{x}) \in \mathbb{R}$).

Proof: Under condition C1, we have

$$R_L(\mathbf{f}) = \mathbb{E}_{\mathbf{x},\mathbf{y_x}}[L(\mathbf{f}(\mathbf{x}), \mathbf{y_x})] = \sum_{q=1}^{k} \mathbb{E}_{\mathbf{x},\mathbf{y_x}} l((\mathbf{f}(\mathbf{x}))_q, (\mathbf{y_x})_q)$$

Since each term in the above sum depends on only one component of $\mathbf{f}$, to minimize R_L we can individually minimize each term and then make the minimizers as components of $\mathbf{f}^*$, the minimizer of risk under noise free case. Thus, we have

$$\mathbb{E}_{\mathbf{x},\mathbf{y_x}} l((\mathbf{f}^*(\mathbf{x}))_q, (\mathbf{y_x})_q) \leq \mathbb{E}_{\mathbf{x},\mathbf{y_x}} l((\mathbf{f}(\mathbf{x}))_q, (\mathbf{y_x})_q),\ \forall \mathbf{f} \qquad (6)$$

Now for the risk under the noisy case, we have

$$\begin{aligned} R_L^\eta(\mathbf{f}) &= \mathbb{E}_{\mathbf{x},\tilde{\mathbf{y}}_\mathbf{x}}[L(\mathbf{f}(\mathbf{x}), \tilde{\mathbf{y}}_\mathbf{x})] = \mathbb{E}_{\mathbf{x},\mathbf{y_x}} \mathbb{E}_{\tilde{\mathbf{y}}_\mathbf{x}|\mathbf{x},\mathbf{y_x}} \sum_{q=1}^{k} l(\mathbf{f}(\mathbf{x}), (\tilde{\mathbf{y}}_\mathbf{x})_q) \\ &= \mathbb{E}_{\mathbf{x},\mathbf{y_x}} \sum_{q=1}^{k} (1-\eta_q) l(\mathbf{f}(\mathbf{x}), (\mathbf{y_x})_q) + \eta_q l(\mathbf{f}(\mathbf{x}), 1 - (\mathbf{y_x})_q), \\ &\qquad \text{using equation (3)} \\ &= \mathbb{E}_{\mathbf{x},\mathbf{y_x}} \sum_{q=1}^{k} (1-\eta_q) l(\mathbf{f}(\mathbf{x}), (\mathbf{y_x})_q) + \eta_q (C - l(\mathbf{f}(\mathbf{x}), (\mathbf{y_x})_q)), \\ &\qquad \text{using condition C2 of the theorem} \\ &= \mathbb{E}_{\mathbf{x},\mathbf{y_x}} \sum_{q=1}^{k} (1-2\eta_q) l(\mathbf{f}(\mathbf{x}), (\mathbf{y_x})_q) + \eta_q C \end{aligned}$$

Recall that $\mathbf{f}^*$ is minimizer of risk under noise-free case. Now we have, for any $\mathbf{f}$,

$$\begin{aligned} R_L^\eta(\mathbf{f}) - R_L^\eta(\mathbf{f}^*) &= \mathbb{E}_{\mathbf{x},\mathbf{y_x}} \left(\sum_{q=1}^{k} (1-2\eta_q) \left[l(\mathbf{f}(\mathbf{x}), (\mathbf{y_x})_q) - l(\mathbf{f}^*(\mathbf{x}), (\mathbf{y_x})_q) \right] \right) \\ &= \sum_{q=1}^{k} (1-2\eta_q) \mathbb{E}_{\mathbf{x},\mathbf{y_x}} \left(l(\mathbf{f}(\mathbf{x}), (\mathbf{y_x})_q) - l(\mathbf{f}^*(\mathbf{x}), (\mathbf{y_x})_q) \right) \\ &\geq 0, \text{ by eq. (6) and } \eta_q < 0.5,\ \forall q \end{aligned}$$

Thus, $\mathbf{f}^*$ also minimizes the risk under noisy distribution and thus completes the proof of the theorem.

REMARK 1. *Theorem 1 is about robustness under symmetric label noise where noise probabilities for different label components can be different. The condition C2 in the theorems is same as the symmetry condition on loss functions which is needed for robustness in multi-class problems. The condition C1 in the theorem says that the loss function for the vector label $\mathbf{y_x}$ can be written as a sum of k loss terms, one for each of the components of $\mathbf{y_x}$ (which is the usual case for multi-label loss functions). In addition, the loss for q^{th} component depends only on q^{th} component of $\mathbf{f}$. This second part is also not a particularly restrictive condition, and the standard Hamming loss satisfies this. This condition is needed only for ensuring $\mathbf{f}^*$ satisfies equation 6. With this assumption (for example, class of functions over which we are minimizing is rich enough to contain the Bayes classifier for each of the underlying binary classification problems), we can have robustness under the more general symmetric label noise also. If assuming that the loss function for q^{th} label component depends only on q^{th} component of $\mathbf{f}$ is restrictive, we can relax this assumption. This is what is done in Theorem-2 below where the condition C2 is appropriately changed. However, in such a case we can prove robustness only under the more restrictive uniformly symmetric noise.*

THEOREM 2. *Multi-label classifier learning under risk minimization is noise-tolerant to uniformly symmetric label noise if $\eta < \frac{1}{2}$ and the loss function, $L(\mathbf{f}(\mathbf{x}), \mathbf{y_x})$, satisfies the following conditions:*

C1. $L(\mathbf{f}(\mathbf{x}), \mathbf{y_x}) = \sum_{q=1}^{k} l(\mathbf{f}(\mathbf{x}), (\mathbf{y_x})_q) \quad \forall \mathbf{x}, \mathbf{y_x}, \mathbf{f}$

C2. $l(\mathbf{f}(\mathbf{x}), 0) + l(\mathbf{f}(\mathbf{x}), 1) = C \quad \forall \mathbf{x}, \mathbf{f}$

Here C is a constant greater than zero and $l(\mathbf{f}(\mathbf{x}), (\mathbf{y_x})_q)$ is a loss function. (Here, $\mathbf{f}(\mathbf{x}) \in \mathbb{R}^k$)

Proof:

$$R_L^\eta(\mathbf{f}) = \mathbb{E}_{\mathbf{x},\tilde{\mathbf{y}}_\mathbf{x}}[L(\mathbf{f}(\mathbf{x}), \tilde{\mathbf{y}}_\mathbf{x})] = \mathbb{E}_{\mathbf{x},\mathbf{y_x}}\mathbb{E}_{\tilde{\mathbf{y}}_\mathbf{x}|\mathbf{x},\mathbf{y_x}} \sum_{q=1}^{k} l(\mathbf{f}(\mathbf{x}), (\tilde{\mathbf{y}}_\mathbf{x})_q)$$

$$= \mathbb{E}_{\mathbf{x},\mathbf{y_x}} \sum_{q=1}^{k} (1-\eta) l(\mathbf{f}(\mathbf{x}), (\mathbf{y_x})_q) + \eta l(\mathbf{f}(\mathbf{x}), 1 - (\mathbf{y_x})_q),$$

using equation (3)

$$= \mathbb{E}_{\mathbf{x},\mathbf{y_x}} \sum_{q=1}^{k} (1-\eta) l(\mathbf{f}(\mathbf{x}), (\mathbf{y_x})_q) + \eta (C - l(\mathbf{f}(\mathbf{x}), (\mathbf{y_x})_q)),$$

using condition C2 of the theorem

$$= (1-2\eta)\mathbb{E}_{\mathbf{x},\mathbf{y_x}} \sum_{q=1}^{k} l(\mathbf{f}(\mathbf{x}), (\mathbf{y_x})_q) + \eta C$$

Hence, for any $\mathbf{f}$ we get

$$R_L^\eta(\mathbf{f}) - R_L^\eta(\mathbf{f}^*) = (1-2\eta)\big(\mathbb{E}_{\mathbf{x},\mathbf{y_x}} \sum_{q=1}^{k} l(\mathbf{f}(\mathbf{x}), (\mathbf{y_x})_q) - \mathbb{E}_{\mathbf{x},\mathbf{y_x}} \sum_{q=1}^{k} l(\mathbf{f}^*(\mathbf{x}), (\mathbf{y_x})_q)\big)$$
$$= (1-2\eta)(R_L(\mathbf{f}) - R_L(\mathbf{f}^*)) \geq 0$$

because $\mathbf{f}^*$ is the minimizer of R_L (cf. equation(2)) and $\eta < 0.5$. As the above is true for any function $\mathbf{f} \in \mathcal{F}$, this implies $\mathbf{f}^*$ will also be minimizer of R_L^η, risk under noisy distribution. This completes the proof of the theorem.

4.1 Loss Functions

Here we consider loss functions that satisfy condition C1 of Theorem 1:

$$L(\mathbf{f}(\mathbf{x}), \mathbf{y}) = \sum_{q=1}^{k} l((\mathbf{f}(\mathbf{x}))_q, (\mathbf{y})_q)$$

Recall that by our definition of a classifier, $\mathbf{f}(\mathbf{x}) \in \mathbb{R}^k$. But from now on we would consider models where $\mathbf{f}(\mathbf{x}) \in [0,\ 1]^k$. This would be the case, for example, if the classifier is a feedforward neural network with sigmoidal output layer.

Different choices for the function l above would result in different loss functions, L. If we take l to be 0–1 loss defined as

$$l(x, y) = \begin{cases} 1 & \text{if } y = 1 \ \&\ x \geq 0.5 \text{ or } y = 0 \ \&\ x < 0.5 \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

then L would be the Hamming loss [36] which is a standard loss function used in multi-label setting. Hamming loss in multi-label setting is analogous to 0–1 loss in single label case.

As it is easy to see, the 0–1 loss defined by equation (7) satisfies the condition C2 of Theorem 1. Hence Hamming loss is noise-tolerant to symmetric label noise in multi-label classification. Hamming loss is not continuous, which makes the risk minimization problem computationally hard.

One can get what may be termed surrogates of Hamming loss by using some other suitable loss for l in place of 0–1 loss. As shown in [9], any classification calibrated binary loss function can be used as l, and risk minimization under the resulting loss function would be consistent. Such a loss function for multi-label problems would also be noise-tolerant for symmetric noise if the binary loss, l, satisfies condition C2 of Theorem 1. Below, we consider three popular binary losses, namely, Binary Cross-Entropy (BCE), Mean Square Error(MSE), and Mean Absolute Error(MAE), for the function l. Of these, only MAE satisfies the condition C2. In our simulation experiments in the next section, we compare the relative noise tolerance of these three losses.

As mentioned earlier, we take $\mathbf{f}(\mathbf{x}) \in [0,\ 1]^k$ because our classifier is a feedforward net with sigmoidal output layer. Thus, we can interpret each component of $\mathbf{f}(\mathbf{x})$ as (posterior) probability for the presence of the corresponding label in the label set. Hence in our definition of the binary losses below, we use p to denote any one component of $\mathbf{f}(\mathbf{x})$. These losses are defined by:

$$l(p, y) = \begin{cases} y log(p) + (1-y) log(1-p) & \text{BCE} \\ (y-p)^2 & \text{MSE} \\ |y-p| & \text{MAE} \end{cases}$$

Now we can calculate the LHS of condition C2 for each of the losses to see whether they satisfy the condition:

$$\sum_{y \in \{0,1\}} l(p, y) = \begin{cases} log(p) + log(1-p) = log(\frac{p}{1-p}) & \text{BCE} \\ (1-p)^2 + p^2 = 2p^2 - 2p + 1 & \text{MSE} \\ |1-p| + |p| = 1 & \text{MAE} \end{cases}$$

As clear from the above, only MAE satisfies C2. The relevant sum is at least bounded in case of MSE while that in the case of BCE is unbounded. Hence using MAE as the binary loss should result in robust learning.

Our theoretical results are about the minimizer of the risk. Ideally, the performance with MAE should not degrade if we are minimizing the actual risk. However, in practice, we can only minimize empirical risk and may not get global minimizer of even empirical risk. Thus, as we see in the simulation section, under empirical risk minimization, the performance of even MAE degrades with label noise, but it degrades less compared to non-symmetric losses.

5 EMPIRICAL RESULTS

In this section, we empirically compare the robustness (to symmetric label noise) of the three losses described earlier. We show results on four multi-label datasets 'scene', 'tmc2007', 'yeast' and 'emotions' [30]. 'scene' [4] is an image dataset, 'tmc2007' [27] is a text dataset(where we used the one with top-500 important attributes), 'emotions' [29] is a music emotion dataset and 'yeast' [7] is a dataset of functional class prediction of genes.

5.1 Simulation Settings

We use feedforward neural networks (NN) as our classifiers. In our simulations, the number of input layer nodes are the same as the number of attributes in a dataset. The number of node in the output layer is the same as the cardinality of the label set. We use sigmoid activation at each of the output nodes. All hidden layers are densely interconnected, and the architecture is different for each dataset. Each of the dense layer outputs is first batch normalized(BN) [13] and then passed through ReLU activation function. We use two dense layers for scene, emotions dataset, and four dense layers for 'tmc' dataset. As there are some correlations among the gene sequences, we use three convolution layers(Conv) followed by a dense layer for yeast dataset. Convolution layer output is only passed through ReLU activation function.

Datasets statistics and the respective hidden layer architectures are given in table 1. In the table, Ca(for Cardinality) is the average number of labels per sample, i.e., $Ca = \frac{1}{N}\sum_{i=1}^{N}|\mathbf{y}_{\mathbf{x}_i}|$ and De(for density) is normalized cardinality, i.e., $De = \frac{1}{k}Ca$.

For training emotions, scene, yeast networks we use RMSprop optimizer (with parameters lr=0.001, rho=0.9, epsilon=1e-08, decay=0.0) and for 'tmc' dataset we use Adam optimizer [16] (learning rate - 0.001, beta1 - 0.9, beta2 - 0.999). We also use dropout of 0.5 after each dense layer (except in 'tmc'). Batch size of 64 is used for scene, yeast, emotions datasets, and 1024 for 'tmc'. For 'tmc', we found that batch normalization alone gives good results for all noise levels with different losses, compared to batch normalization and dropout regularizers used together. Hence we used only BN for 'tmc'. For training the networks for 'scene', 'yeast' and 'emotions' datasets, we also reduce learning rate when there is no improvement (i.e., it is ≤ 0.0001) by a factor of 0.1 with the minimum rate set to 10^{-5} and patience 10. We set aside 25% of the noisy training dataset for validation.[1] We train each network for 500 epochs and choose the best network based on the error on the validation data.

5.2 Results

We show results with MAE, MSE, BCE losses and at noise rates – 0%, 10%, 20%, 30%. Symmetric label noise is introduced in the training and validation sets by independently corrupting each label of each sample. We report results on test set which has no label noise. For each dataset, loss and noise rate, we train the network five times to account for randomness in label noise and initial starting point. We report results using the commonly used evaluation metrics for multi-label classifier learning – F1-macro, F1-micro and Jaccard Index [36].

$$F1_{macro} = \frac{1}{k}\sum_{q=1}^{k}\frac{2*TP_q}{2*TP_q+FN_q+FP_q}$$

$$F1_{micro} = \frac{2*\sum_{q=1}^{k}TP_q}{2*\sum_{q=1}^{k}TP_q+\sum_{q=1}^{k}FN_q+\sum_{q=1}^{k}FP_q}$$

$$Jaccard\ Index = \sum_{i=1}^{N}\frac{|y_{\mathbf{x}_i}\wedge\mathbf{h}(\mathbf{x}_i)|}{|y_{\mathbf{x}_i}\vee\mathbf{h}(\mathbf{x}_i)|}$$

where TP, FP and FN are label wise true positives, false positives, false negatives respectively. $\wedge$ and $\vee$ are element wise logical AND and logical OR operations. (Here $|s|$ represents number of ones in the binary vector s).

The results are shown in Figure 1 for the four datasets. Each figure shows the values of F1-macro, F1-micro, and Jaccard Index for different noise rates. (We show the mean and standard deviation). As can be seen, on all the three metrics, the MAE loss shows better robustness compared to the other two. (Recall that MAE is the only loss that satisfies the symmetry condition of our theorem). This robustness is particularly good for the 'yeast' and 'emotions' datasets. Ideally, the performance with MAE should not degrade if we are minimizing the true risk. If we could find the global minimizer of true risk, then the performance of MAE should be the same with or without noise. However, here we are only minimizing empirical risk, and the neural network algorithms do not necessarily learn the global minimum even of the empirical risk. Also, normally, sample complexities under label noise are higher, and the small number of examples may be another reason why MAE performance also degrades. However, MAE has much less degradation compared to the other two losses. This adequately illustrates the utility of our theoretical results.

The point that we wish to make, based on simulations, is that using a robust loss function is attractive even if we do not get the ideal robustness (because we are only finding local minima of empirical risk). This is because, with a robust loss function, one gains some robustness without any extra cost. If the data does not have noise, the performance of MAE is on par with that of BCE; under noise, the performance of MAE is better. For the learning algorithm, we need no information about the noise (not even any information at all about whether there is label noise). We use a different loss function in empirical risk minimization. In this sense whatever robustness we are getting is for free. This is what is attractive about robust loss functions.

6 CONCLUSION

In this paper, we considered learning of multi-label classifiers in presence of label noise. We defined symmetric label noise model in a multi-label setting which can model random errors by labelers. Random label noise in the multi-label setting is unavoidable because of the inherent uncertainty (or variability in the decisions of different labelers) in deciding which subset of labels is relevant for a given instance. Thus, it is essential to consider classifier learning strategies that are robust to random label noise.

[1] Noisy data can be used for validation as the risk under noisy distribution is linearly related to risk under noise-free distribution when the loss function satisfies the sufficient conditions in Theorem 2.

Table 1: Datasets and Architecture. Abbreviations:- DL(integer) - Dense layer(hidden units), BN - Batch Normalization, AF - activation function layer, dr - dropout), Conv(filters) - Convolution Layer(filters). We use ReLU activation function everywhere.

Dataset	instances	attributes	labels	Ca/De	Hidden layer Architecture
scene	2407	294	6	1.074/0.179	DL(64) + BN + AF + (dr = 0.5)+ DL(32) + BN + AF + (dr = 0.5)
tmc2007	28596	500	22	2.158/0.098	DL(1024)+BN+AF+DL(512)+BN+AF+ DL(256)+BN+AF + DL(32)+BN+AF
yeast	2417	103	14	4.237/0.303	Conv(8)+ AF + Conv(16)+ AF+ Conv(32)+ AF+ DL(16) + BN + AF + (dr = 0.5)
emotions	593	72	6	1.869/0.311	DL(64) + BN + AF + (dr = 0.5) + DL(16) + BN + AF + (dr = 0.5)

The focus of this paper is on the robustness of risk minimization in multi-label settings. Many classifier learning algorithms can be cast in the framework of risk minimization. In this paper, we have proved a sufficient condition on the loss function for risk minimization to be robust to symmetric label noise. Hamming loss, which uses 0–1 loss as the single-label loss, satisfies this sufficient condition. We also show that a surrogate of Hamming loss obtained by replacing 0–1 loss with MAE loss is tolerant to symmetric label noise. On the other hand, other losses, such as BCE or MSE, are not robust. Through empirical results on benchmark datasets, we illustrated this robustness. Our simulation results clearly show the superiority of MAE over other losses. We do not compare the robustness of other popular multi-label learning algorithms like Multi-label Decision Tree, which can not be represented as risk minimization.

Our theoretical results pertain to the minimization of true risk, though in practice, one can minimize only the empirical risk. Hence, a useful extension of results presented here would be in the direction of bounds on generalization error for empirical risk minimization in the presence of label noise. In the presence of label noise, one would need more examples for consistent learning, and such bounds would be useful for adequately exploiting our results on robustness. In this paper, we consider symmetric label noise in which for any label, the probability of it being wrongly present is the same as it being wrongly absent. (Even though this probability can be different for different labels). Many scenarios (other than labeling errors) such as missing labels, partially labeled data, etc. can be modeled using label noise. Many such situations may not satisfy our symmetric noise assumption. A more challenging problem would be to consider the non-symmetric case or the general case where the noise probability could be a function of the feature vector. This is also a very promising direction in which results presented here need to be extended.

REFERENCES

[1] Anelia Angelova, Yaser Abu-Mostafa, and Pietro Perona. 2005. Pruning Training Sets for Learning of Object Categories. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Washington, DC, USA, 494–501.

[2] Blaine Nelson Battista Biggio and Pavel Laskov. 2011. Support Vector Machines Under Adversarial Label Noise. In *Proceedings of the Third Asian Conference on Machine Learning*. Taoyuan, Taiwan, 97–112.

[3] Jakramate Bootkrajang and Ata Kabán. 2012. Label-noise robust logistic regression and its applications. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 143–158.

[4] Matthew R. Boutell, Jiebo Luo, Xipeng Shen, and Christopher M. Brown. 2004. Learning multi-label scene classification. *Pattern Recognition* 37, 9 (2004), 1757 – 1771. https://doi.org/10.1016/j.patcog.2004.03.009

[5] Carla E. Brodley and Mark A. Friedl. 1999. Identifying Mislabeled Training Data. *Journal Of Artificial Intelligence Research* 11 (August 1999), 131–167.

[6] B. Du, Z. Wang, L. Zhang, L. Zhang, and D. Tao. 2017. Robust and Discriminative Labeling for Multi-Label Active Learning Based on Maximum Correntropy Criterion. *IEEE Transactions on Image Processing* 26, 4 (April 2017), 1694–1707. https://doi.org/10.1109/TIP.2017.2651372

[7] André Elisseeff and Jason Weston. 2002. A kernel method for multi-labelled classification. In *Advances in neural information processing systems*. 681–687.

[8] Benoît Frénay and Michel Verleysen. 2014. Classification in the Presence of Label Noise: A Survey. *IEEE Transactions on Neural Networks and Learning Systems* 25, 5 (May 2014), 845–869.

[9] Wei Gao and Zhi-Hua Zhou. 2011. On the consistency of multi-label learning. In *Proceedings of the 24th annual conference on learning theory*. 341–358.

[10] Aritra Ghosh, Himanshu Kumar, and P. S. Sastry. 2017. Robust Loss Functions under Label Noise for Deep Neural Networks. (2017). https://www.aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14759

[11] Aritra Ghosh, Naresh Manwani, and PS Sastry. 2015. Making risk minimization tolerant to label noise. *Neurocomputing* 160 (2015), 93–107.

[12] Mengying Hu, Hu Han, Shiguang Shan, and Xilin Chen. 2019. Multi-label Learning from Noisy Labels with Non-linear Feature Transformation. In *Computer Vision – ACCV 2018*. 404–419.

[13] Sergey Ioffe and Christian Szegedy. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *CoRR* abs/1502.03167 (2015). arXiv:1502.03167 http://arxiv.org/abs/1502.03167

[14] Rong Jin, Yan Liu, Luo Si, Jaime G Carbonell, and Alexander Hauptmann. 2003. A new boosting algorithm using input-dependent regularizer. In *Proceedings of Twentieth International Conference on Machine Learning*. Washington D.C.

[15] Roni Khardon and Gabriel Wachman. 2007. Noise Tolerant Variants of the Perceptron Algorithm. *Journal Of Machine Learning Research* 8 (February 2007), 227–248.

[16] Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *CoRR* abs/1412.6980 (2014). arXiv:1412.6980 http://arxiv.org/abs/1412.6980

[17] Neil D Lawrence and Bernhard Schölkopf. 2001. Estimating a kernel Fisher discriminant in the presence of label noise. In *ICML*, Vol. 1. Citeseer, 306–313.

[18] Philip M Long and Rocco A Servedio. 2010. Random classification noise defeats all convex potential boosters. *Machine Learning* 78, 3 (2010), 287–304.

[19] Naresh Manwani and PS Sastry. 2013. Noise tolerance under risk minimization. *Cybernetics, IEEE Transactions on* 43, 3 (2013), 1146–1151.

[20] V. Mnih and G. E. Hinton. 2012. Learning to Label Aerial Images from Noisy Data. In *ICML*. Citeseer.

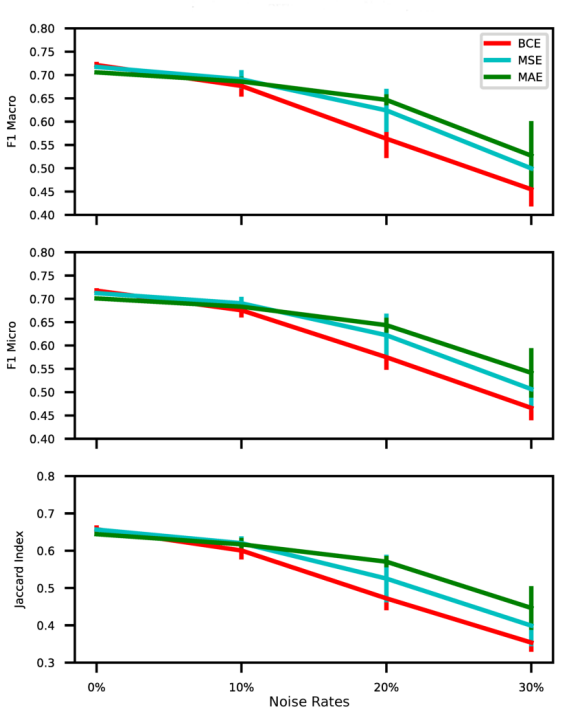
[21] Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. 2013. Learning with noisy labels. In *Advances in neural information processing systems*. 1196–1204.

[22] Divya Padmanabhan, Satyanath Bhat, Shirish Shevade, and Y. Narahari. 2017. Multi-Label Classification from Multiple Noisy Sources Using Topic Models. *Information* 8, 2 (2017).

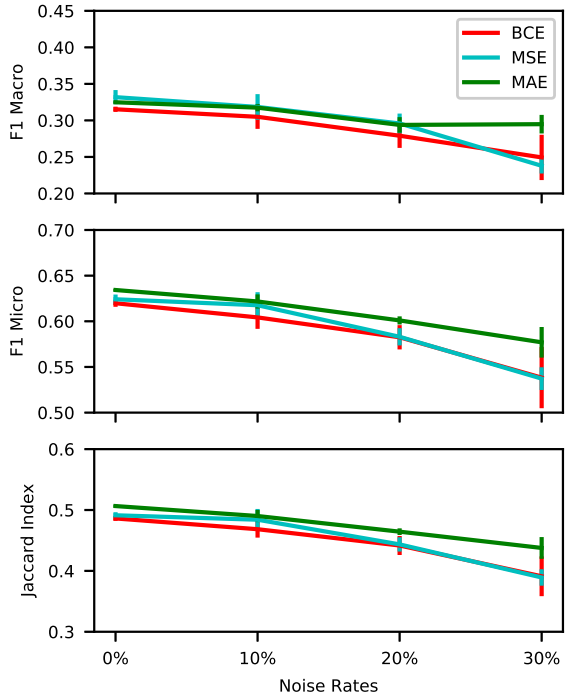
[23] Giorgio Patrini, Alessandro Rozza, Aditya Menon, Richard Nock, and Lizhen Qu. 2017. Making neural networks robust to label noise: a loss correction approach. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[24] Scott Reed, Honglak Lee, Dragomir Anguelov, Christian Szegedy, Dumitru Erhan, and Andrew Rabinovich. 2014. Training deep neural networks on noisy labels with bootstrapping. *arXiv preprint arXiv:1412.6596* (2014).

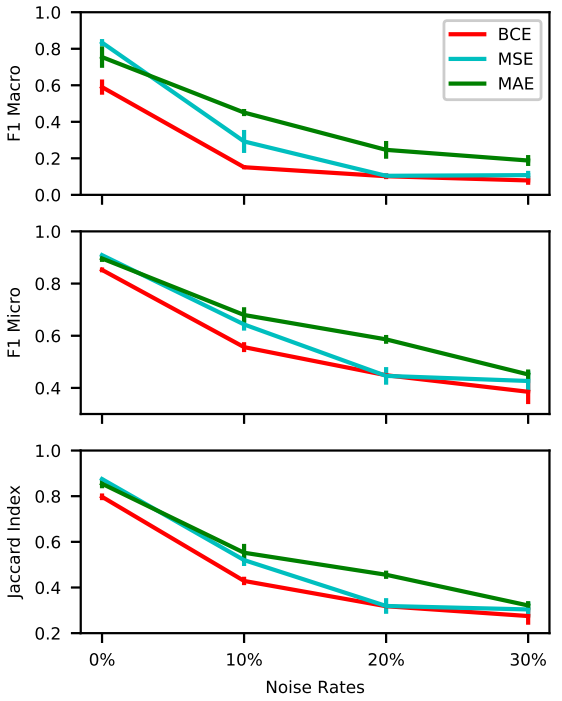
[25] PS Sastry and Naresh Manwani. 2016. Robust Learning of Classifiers in the Presence of Label Noise. *Pattern Recognition and Big Data* (2016), 167.



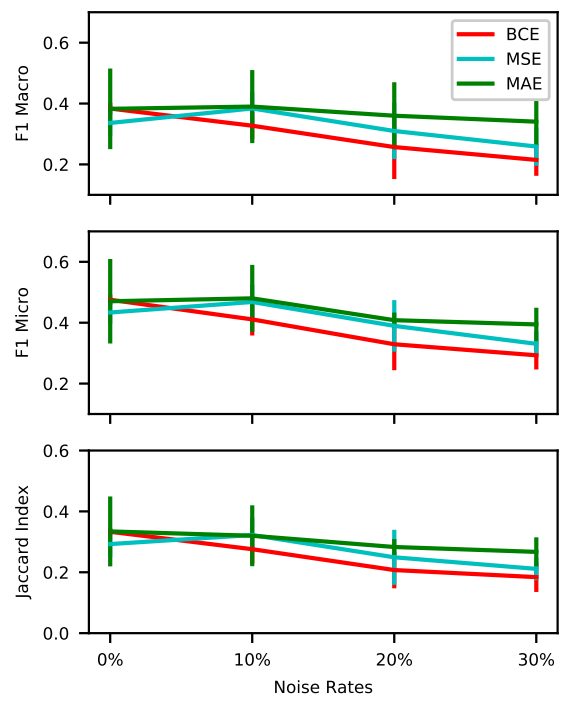
(a) Scene Dataset



(b) Yeast Dataset



(c) tmc500 Dataset



(d) emotion Dataset

Figure 1: F1 Scores for different datasets.

[26] Clayton Scott, Gilles Blanchard, and Gregory Handy. 2013. Classification with Asymmetric Label Noise: Consistency and Maximal Denoising. In *COLT 2013 - The 26th Annual Conference on Learning Theory, June 12-14, 2013, Princeton University, NJ, USA*. 489–511.

[27] A. N. Srivastava and B. Zane-Ulman. 2005. Discovering recurring anomalies in text reports regarding complex space systems. In *2005 IEEE Aerospace Conference*. 3853–3862. https://doi.org/10.1109/AERO.2005.1559692

[28] Sainbayar Sukhbaatar and Rob Fergus. 2014. Learning from Noisy Labels with Deep Neural Networks. *CoRR* abs/1406.2080 (2014). arXiv:1406.2080 http://arxiv.org/abs/1406.2080

[29] Konstantinos Trohidis, Grigorios Tsoumakas, George Kalliris, and Ioannis Vlahavas. 2011. Multi-label classification of music by emotion. *EURASIP Journal on Audio, Speech, and Music Processing* 2011, 1 (18 Sep 2011), 4. https://doi.org/10.1186/1687-4722-2011-426793

[30] Grigorios Tsoumakas, Eleftherios Spyromitros-Xioufis, Jozef Vilcek, and Ioannis Vlahavas. 2011. MULAN: A Java Library for Multi-Label Learning. *J. Mach. Learn. Res.* 12 (July 2011), 2411–2414. http://dl.acm.org/citation.cfm?id=1953048.2021078

[31] Naonori Ueda and Kazumi Saito. 2002. Parametric Mixture Models for Multi-labeled Text. In *Proceedings of the 15th International Conference on Neural Information Processing Systems (NIPS'02)*. MIT Press, Cambridge, MA, USA, 737–744. http://dl.acm.org/citation.cfm?id=2968618.2968710

[32] Brendan van Rooyen, Aditya Menon, and Robert C Williamson. 2015. Learning with symmetric label noise: The importance of being unhinged. In *Advances in Neural Information Processing Systems*. 10–18.

[33] Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. 2015. Learning From Massive Noisy Labeled Data for Image Classification. In *The IEEE Conference on Computer Vision and Pattern Recognition*.

[34] Chang Xu*, Dacheng Tao, and Chao Xu. 2016. Robust Extreme Multi-label Learning. *ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (August 2016).

[35] Min-Ling Zhang and Zhi-Hua Zhou. 2006. Multilabel Neural Networks with Applications to Functional Genomics and Text Categorization. *IEEE Transactions on Knowledge and Data Engineering* 18, 10 (Oct 2006), 1338–1351. https://doi.org/10.1109/TKDE.2006.162

[36] M. L. Zhang and Z. H. Zhou. 2014. A Review on Multi-Label Learning Algorithms. *IEEE Transactions on Knowledge and Data Engineering* 26, 8 (Aug 2014), 1819–1837. https://doi.org/10.1109/TKDE.2013.39

[37] Xingquan Zhu, Xindong Wu, and Qijun Chen. 2003. Eliminating class noise in large datasets. In *Proceedings of the Twentieth International Conference on Machine Learning*. Washington, DC, USA, 920–927.