Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)

Robust Loss Functions under Label Noise for Deep Neural Networks

Aritra Ghosh
arghosh@microsoft.com
Microsoft, Bangalore

Himanshu Kumar
himanshukr@ee.iisc.ernet.in
Indian Institute of Science, Bangalore

P. S. Sastry
sastry@ee.iisc.ernet.in
Indian Institute of Science, Bangalore

Abstract

In many applications of classifier learning, training data suffers from label noise. Deep networks are learned using huge training data where the problem of noisy labels is particularly relevant. The current techniques proposed for learning deep networks under label noise focus on modifying the network architecture and on algorithms for estimating true labels from noisy labels. An alternate approach would be to look for loss functions that are inherently noise-tolerant. For binary classification there exist theoretical results on loss functions that are robust to label noise. In this paper, we provide some sufficient conditions on a loss function so that risk minimization under that loss function would be inherently tolerant to label noise for multiclass classification problems. These results generalize the existing results on noise-tolerant loss functions for binary classification. We study some of the widely used loss functions in deep networks and show that the loss function based on mean absolute value of error is inherently robust to label noise. Thus standard back propagation is enough to learn the true classifier even under label noise. Through experiments, we illustrate the robustness of risk minimization with such loss functions for learning neural networks.

Introduction

Recently, deep neural networks have exhibited very impressive performance in many classification problems. However, in all such cases one needs very large training data. Labelling the training data patterns and ensuring correctness of the labels thus becomes a serious challenge in many applications of deep neural networks.

When the class labels in the training data are noisy (i.e., may be incorrect) then it is referred to as label noise. Human labelling errors, measurement errors, subjective biases of labellers are among the reasons for label noise. In many large scale classification problems, labelled data is often obtained through crowd sourcing or by automatically using information on the web. This is another main reason for unreliability of labels in the training data.

Robust learning of classifiers in the presence of label noise has been investigated from many viewpoints. In this paper, we study it in the framework of risk minimization which is a popular method for classifier learning. For example, Bayes classifier minimizes risk under 0–1 loss function. The standard backpropagation-based learning of neural networks is also risk minimization under different loss functions (such as squared error or cross entropy). The robustness of risk minimization depends on the loss function used. We call a loss function noise-tolerant if minimizer of risk (under that loss function) with noisy labels would be same as that with noise-free labels.

In this paper we present some novel analytical results on noise-tolerance of loss functions in a multiclass classification scenario. We derive sufficient conditions on a loss function so that it would be noise-tolerant for different types of label noise. We then examine some of the popular loss functions used for learning neural networks and show that loss function based on mean-absolute error (MAE) satisfies our sufficient conditions. Empirical investigations are presented to compare robustness of learning neural networks under label noise using different loss functions based on mean-absolute error, mean-square error and categorical cross entropy. The empirical results well demonstrate the utility of the theoretical results presented here.

Related Work

Learning in presence of label noise is a long standing problem in machine learning. A detailed survey is available in (Frénay and Verleysen 2014).

There are many approaches to learning under label noise. Data cleaning approaches rely on finding points which are corrupted by label noise. Once these points are identified, they can be either filtered out or their labels suitably altered. Several heuristics have been used to guess such noisy points (Angelova, Abu-Mostafa, and Perona 2005; Brodley and Friedl 1999; Zhu, Wu, and Chen 2003). There have also been attempts at (heuristically) modifying existing learning algorithms to make them robust (Khardon and Wachman 2007; Jin et al. 2003; Battista Biggio and Laskov 2011).

Another prominent approach is to treat the (unknown) true labels as hidden variables and to estimate a generative or discriminative model. Lawrence and Schölkopf proposed such a method, based on maximum-likelihood estimation of the model using the EM algorithm, in the context of fisher linear discriminant classifier. Similar methods have been proposed to make logistic regression robust to label noise (Bootkra-

jang and Kabán 2012). Recently many algorithms based on such ideas are proposed in deep learning literature to mitigate the adverse effect of label noise. Mnih and Hinton use a generative model for label corruption, estimated through an approximate EM algorithm, and show its effectiveness in a binary classification problem. Sukhbaatar et al. proposed a modified architecture of a neural network to learn with noisy labels by effectively estimating label corruption probabilities. Motivated by similar ideas, a method to estimate a generative model incorporating label corruption is proposed in (Xiao et al. 2015). This method is applicable for multiclass classification and can handle fairly general cases of label noise. Methods based on bootstrapping are also proposed for learning deep networks under label noise (Reed et al. 2014). While all these methods are seen to deliver good performance, they do not guarantee, in any probabilistic sense, robustness to label noise.

All the above methods focus on changing the learning algorithm so that one can estimate the true labels of the training examples and thus be able to learn under label noise. As opposed to this, one can also look for methods that are inherently noise tolerant. Such algorithms treat noisy data and noise-free data the same way but achieve noise robustness due to properties of the algorithm. Such methods have been mostly investigated in the framework of risk minimization.

Robustness of risk minimization depends on the loss function. For binary classification, it is shown that 0–1 loss is robust to symmetric or uniform label noise while most of the standard convex loss functions are not robust (Long and Servedio 2010; Manwani and Sastry 2013). The problem of learning from positive and unlabeled data can be cast as learning under label noise and it is seen that none of the common convex surrogate losses are good for this problem (du Plessis, Niu, and Sugiyama 2014). Unhinged loss, a convex loss (which is not convex potential), is robust to symmetric label noise (van Rooyen, Menon, and Williamson 2015). Natarajan et al. proposed a method for robust risk minimization through an implicit estimation of noise probabilities. In a similar spirit, Scott, Blanchard, and Handy proposed a method of estimating Type 1 and Type 2 error rates of any specific classifier under the noise-free distribution given only the noisy training data. Recently, a general sufficient condition on a loss function is derived so that risk minimization is robust to label noise (Ghosh, Manwani, and Sastry 2015). It is shown that the 0–1 loss, ramp loss and sigmoidal loss satisfy this condition.

All the above are for the case of binary classification. Recently Patrini et al. proposed a robust risk minimization approach for learning neural networks for multiclass classification by estimating label corruption probabilities. In our work here we also investigate robustness of risk minimization in the context of multiclass classification. We provide analytical results on conditions under which risk minimization is robust to different types of label noise. Our results generalize the existing results for 2-class problems. The currently known noise-tolerant loss functions (such as 0–1 loss or ramp loss) are not commonly used while learning neural networks. In this paper, we examine some common loss functions for learning neural networks for multiclass classification and show that the one based on mean absolute error is noise-tolerant. Through empirical studies we demonstrate the relevance of our theoretical results.

Preliminaries and Problem Statement

In this section we introduce some notation and define the notion of noise tolerance of a loss function.

Risk Minimization

Let $\mathcal{X} \subset \mathbb{R}^d$ be the feature space from which the examples are drawn and let $\mathcal{Y} = [k] = \{1, \cdots, k\}$ be the class labels. In a typical classifier learning problem, we are given training data, $S = \{(\mathbf{x}_1, y_{\mathbf{x}_1}), \ldots, (\mathbf{x}_N, y_{\mathbf{x}_N})\} \in (\mathcal{X} \times \mathcal{Y})^N$, drawn *iid* according to an unknown distribution, $\mathcal{D}$, over $\mathcal{X} \times \mathcal{Y}$. We represent a classifier as $h(\mathbf{x}) = \text{pred} \circ f(\mathbf{x})$ where $f : \mathcal{X} \to \mathcal{C},\ \mathcal{C} \subseteq \mathbb{R}^k$. Here, h (which predicts the class label given $f(\mathbf{x})$) maps $\mathcal{X}$ to $\mathcal{Y}$. Even though the final classification decision on a feature vector $\mathbf{x}$ is $\text{pred} \circ f(\mathbf{x})$, we use the notation of calling f itself as the classifier.

A loss function is a map $L : \mathcal{C} \times \mathcal{Y} \to \mathbb{R}^+$. Given any loss function, L, and a classifier, f, we define the *L-risk* of f by

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

where, as a notation throughout this paper, the $\mathbb{E}$ denotes expectation and its subscript indicates the random variables or the distribution with respect to which the expectation is taken. Under risk minimization framework, the objective is to learn a classifier, f, which is a global minimizer of R_L. Note that the L-risk, R_L, depends on L, the loss function. When L happens to be the 0–1 loss, R_L would be the usual Bayes risk.

Noise Tolerance of Loss Functions

When there is label noise, the learner does not have access to the clean training data (represented by S above). The noisy training data available to the learner is $S_\eta = \{(\mathbf{x}_n, \hat{y}_{\mathbf{x}_n}), n = 1, \cdots, N\}$ where,

$$\hat{y}_{\mathbf{x}_n} = \begin{cases} y_{\mathbf{x}_n} & \text{with probability } (1 - \eta_{\mathbf{x}_n}) \\ j,\ \ j \in [k],\ j \neq y_{\mathbf{x}_n} & \text{with probability } \bar{\eta}_{\mathbf{x}_n j} \end{cases}$$

Note that, for all $\mathbf{x}$, conditioned on $y_{\mathbf{x}} = i$, we have $\sum_{j \neq i} \bar{\eta}_{\mathbf{x}j} = \eta_{\mathbf{x}}$.

In general, for any $\mathbf{x}$, its true label (that is, label under distribution $\mathcal{D}$) is denoted by the random variable $y_{\mathbf{x}}$ while the noise corrupted label is denoted by $\hat{y}_{\mathbf{x}}$. We use $\mathcal{D}_\eta$ to denote the joint probability distribution of $\mathbf{x}$ and $\hat{y}_{\mathbf{x}}$.

The noise is termed *symmetric* or *uniform* if $\eta_{\mathbf{x}} = \eta$, and $\bar{\eta}_{\mathbf{x}j} = \frac{\eta}{k-1},\ \forall j \neq y_{\mathbf{x}}, \forall \mathbf{x}$, where η is a constant.

Noise is said to be *class-conditional* or asymmetric if the dependence of $\eta_{\mathbf{x}}$ on $\mathbf{x}$ is only through $y_{\mathbf{x}}$ and similarly for $\bar{\eta}_{\mathbf{x}j}$. In this case, with a little abuse of notation, we write $\eta_{\mathbf{x}} = \eta_{y_{\mathbf{x}}},\ \bar{\eta}_{\mathbf{x}j} = \bar{\eta}_{y_{\mathbf{x}}j}$. Thus, for example, $\bar{\eta}_{ij}$ would be the probability that a class-i pattern would have label as class-j when the label is corrupted.

In general, when noise rate $\eta_{\mathbf{x}}$ as well as $\bar{\eta}_{\mathbf{x}j}$ is a function of $\mathbf{x}$, it is termed as *non-uniform* noise. A simple special case is when $\bar{\eta}_{\mathbf{x}j} = \frac{\eta_{\mathbf{x}}}{k-1},\ \forall j \neq y_{\mathbf{x}}$. We define it as *simple*

non-uniform noise. Furthermore, when $\eta_{\mathbf{x}}$ is fixed for each class, we call it *simple class conditional noise*.

The L-risk, $R_L(f)$, given by eq.1 is for the noise-free case. Let f^* be the global minimizer (over the chosen function class) of $R_L(f)$. When there is label noise, the data is drawn according to distribution $\mathcal{D}_\eta$. Then L-risk of a classifier f under noisy data is

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

(We use $\mathbb{E}_{\mathbf{x},y_{\mathbf{x}}}(\mathbb{E}_{\mathbf{x},\hat{y}_{\mathbf{x}}})$ and $\mathbb{E}_{\mathcal{D}}(\mathbb{E}_{\mathcal{D}_\eta})$ interchangeably). Let f_η^* be the global minimizer (over the chosen function class) of $R_L^\eta(f)$. Risk minimization under loss function L, is said to be *noise-tolerant* if (Manwani and Sastry 2013)

$$\Pr_{\mathcal{D}}[\text{pred} \circ f^*(\mathbf{x}) = y_{\mathbf{x}}] = \Pr_{\mathcal{D}}[\text{pred} \circ f_\eta^*(\mathbf{x}) = y_{\mathbf{x}}]$$

Risk minimization under a given loss function is noise tolerant if the f_η^* has the same probability of misclassification as that of f^* on the noise free data. When the above is satisfied we also say that the loss function L is noise-tolerant. For this, it is sufficient if $f^* = f_\eta^*$.

Theoretical Results

We call a loss function L symmetric if it satisfies, for some constant C,

$$\sum_{i=1}^{k} L(f(\mathbf{x}), i) = C, \ \forall \mathbf{x} \in \mathcal{X}, \forall f. \tag{2}$$

In the following, we prove distribution independent sufficient conditions for loss function to be robust under different kinds of label noises. In Theorem 1, we prove sufficiency results for *symmetric label noise*, followed by *simple non-uniform noise* and *class-conditional noise* in Theorems 2, 3.

Theorem 1 *In a multi-class classification problem, let loss function L satisfy Eq 2. Then L is noise tolerant under symmetric or uniform label noise if $\eta < \frac{k-1}{k}$.*

Proof 1 *Recall that for any f,*

$$R_L(f) = \mathbb{E}_{\mathbf{x},y_{\mathbf{x}}} L(f(\mathbf{x}), y_{\mathbf{x}}) \tag{3}$$

For uniform noise, we have, for any f,[1]

$$\begin{aligned} R_L^\eta(f) &= \mathbb{E}_{\mathbf{x},\hat{y}_{\mathbf{x}}} L(f(\mathbf{x}), \hat{y}_{\mathbf{x}}) \\ &= \mathbb{E}_{\mathbf{x}} \mathbb{E}_{y_{\mathbf{x}}|\mathbf{x}} \mathbb{E}_{\hat{y}_{\mathbf{x}}|\mathbf{x},y_{\mathbf{x}}} L(f(\mathbf{x}), \hat{y}_{\mathbf{x}}) \\ &= \mathbb{E}_{\mathbf{x}} \mathbb{E}_{y_{\mathbf{x}}|\mathbf{x}} \left[(1-\eta) L(f(\mathbf{x}), y_{\mathbf{x}}) + \frac{\eta}{k-1} \sum_{i \neq y_{\mathbf{x}}} L(f(\mathbf{x}), i) \right] \\ &= (1-\eta) R_L(f) + \frac{\eta}{k-1}(C - R_L(f)) \\ &= \frac{C\eta}{k-1} + \left(1 - \frac{\eta k}{k-1}\right) R_L(f). \end{aligned}$$

Thus, for any f,

$$R_L^\eta(f^*) - R_L^\eta(f) = (1 - \frac{\eta k}{k-1})(R_L(f^*) - R_L(f)) \leq 0$$

because $\eta < \frac{k-1}{k}$ and f^* is a minimizer of R_L. This proves f^* is also minimizer of risk under uniform noise.

[1] In the following, $\mathbb{E}_{y_{\mathbf{x}}|\mathbf{x}} \mathbb{E}_{\hat{y}_{\mathbf{x}}|\mathbf{x},y_{\mathbf{x}}}$ etc. denote expectation with respect to the corresponding conditional distributions. Note that $\mathbb{E}_{\mathbf{x}} \mathbb{E}_{y_{\mathbf{x}}|\mathbf{x}} = \mathbb{E}_{\mathbf{x},y_{\mathbf{x}}} = \mathbb{E}_{\mathcal{D}}$.

Remark 1 Theorem 1 shows that symmetric losses are robust to uniform label noise. This does not depend on the data distribution. The only condition is that noise rate is less than $\frac{k-1}{k}$ which is not restrictive. This theorem (along with the next one) generalizes the existing results for the 2-class case (Ghosh, Manwani, and Sastry 2015, Theorem 1).

Theorem 2 *Suppose loss L satisfies Eq 2. If $R_L(f^*) = 0$, then L is also noise tolerant under simple non uniform noise when $\eta_{\mathbf{x}} < \frac{k-1}{k}$, $\forall \mathbf{x}$.*
If $R_L(f^*) = \rho > 0$ then, under simple non-uniform noise, $R_L(f_\eta^*)$ is upper bounded by $\rho/(1 - \frac{k\eta_{max}}{k-1})$, where η_{max} is maximum noise rate over $\mathbf{x} \in \mathcal{X}$. (Recall that f^* is minimizer of R_L and f_η^* is minimizer of R_L^η).

Proof 2 *Under simple non-uniform noise, for any f,*

$$\begin{aligned} R_L^\eta(f) &= \mathbb{E}_{\mathbf{x}} \mathbb{E}_{y_{\mathbf{x}}|\mathbf{x}} \mathbb{E}_{\hat{y}_{\mathbf{x}}|\mathbf{x},y_{\mathbf{x}}} L(f(\mathbf{x}), \hat{y}_{\mathbf{x}}) \\ &= \mathbb{E}_{\mathcal{D}} \left[(1-\eta_{\mathbf{x}}) L(f(\mathbf{x}), y_{\mathbf{x}}) + \sum_{i \neq y_{\mathbf{x}}} \frac{\eta_{\mathbf{x}} L(f(\mathbf{x}), i)}{k-1} \right] \\ &= \mathbb{E}_{\mathcal{D}} (1-\eta_{\mathbf{x}}) L(f(\mathbf{x}), y_{\mathbf{x}}) \\ &\quad + \mathbb{E}_{\mathcal{D}} \frac{\eta_{\mathbf{x}}}{k-1} \Big(C - L(f(\mathbf{x}), y_{\mathbf{x}}) \Big) \\ &= \mathbb{E}_{\mathcal{D}} C' \eta_{\mathbf{x}} + \mathbb{E}_{\mathcal{D}} \left((1 - \frac{k\eta_{\mathbf{x}}}{k-1}) L(f(\mathbf{x}), y_{\mathbf{x}}) \right) \end{aligned}$$

where $C' = \frac{C}{k-1}$. Hence we have

$$\begin{aligned} R_L^\eta(f^*) - R_L^\eta(f) = &\mathbb{E}_{\mathcal{D}} \big\{ (1 - \frac{k\eta_{\mathbf{x}}}{k-1}) \\ &(L(f^*(\mathbf{x}), y_{\mathbf{x}}) - L(f(\mathbf{x}), y_{\mathbf{x}})) \big\} \end{aligned} \tag{4}$$

Since $R_L(f^*) = 0$ and L is non-negative by definition, we have $L(f^*(\mathbf{x}), y_{\mathbf{x}}) = 0, \ \forall \mathbf{x}$. In addition, since $(1 - \frac{k\eta_{\mathbf{x}}}{k-1}) > 0$, we have $R_L^\eta(f^*) - R_L^\eta(f) \leq 0$, for any f. Thus minimizer of noise free case is also a minimizer of noisy case. This completes proof of first part of theorem.

For the second part of the theorem, we have,

$$\begin{aligned} & R_L^\eta(f_\eta^*) - R_L^\eta(f^*) \leq 0 \\ \Rightarrow & \mathbb{E}_{\mathcal{D}} (1 - \frac{k\eta_{\mathbf{x}}}{k-1})(L(f_\eta^*(\mathbf{x}), y_{\mathbf{x}}) - L(f^*(\mathbf{x}), y_{\mathbf{x}})) \leq 0 \\ \Rightarrow & \min_{\eta_{\mathbf{x}}} (1 - \frac{k\eta_{\mathbf{x}}}{k-1}) \mathbb{E}_{\mathcal{D}} L(f_\eta^*(\mathbf{x}), y_{\mathbf{x}}) \leq \mathbb{E}_{\mathcal{D}} L(f^*(\mathbf{x}), y_{\mathbf{x}}) \\ & \Rightarrow R_L(f_\eta^*) \leq \rho/(1 - \frac{k\eta_{max}}{k-1}) \end{aligned} \tag{5}$$

where $\mathbb{E}_{\mathcal{D}} L(f^*(\mathbf{x}), y_{\mathbf{x}}) = R_L(f^*) = \rho$. Note that, in the above, we used $\eta_{\mathbf{x}} < \frac{k-1}{k}$, and hence $0 < (1 - \frac{k\eta_{\mathbf{x}}}{k-1}) \leq 1, \ \forall \mathbf{x}$. This completes the proof.

Remark 2 Theorem 2 establishes a sufficient condition for risk minimization to be robust to simple non uniform label noise. The condition needs $R_L(f^*)$ to be zero. If L is the 0–1 loss then R_L is the Bayes risk and then the sufficient condition is that the classes are separable (under noise-free case). However, even if the classes are separable, for a general loss

function (e.g., sigmoidal loss), $R_L(f^*)$ may not be zero. The second part of Theorem 2 gives a bound on $R_L(f^*_\eta)$ in such cases. This part is useful even when classes are not separable and the optimal Bayes risk is non-zero. In case of high noise rate the bound might be loose, but one should note that, this is a distribution independent bound. In the binary classification case, if data is separable, robustness can be achieved even though minimum value of L-risk is not zero if the loss function is 'sufficiently steep' (Ghosh, Manwani, and Sastry 2015, Theorems 2, 4). It appears possible to prove a similar result in multiclass case also.

Theorem 3 *Suppose L satisfies Eq 2 and $0 \leq L(f(\mathbf{x}), i) \leq C/(k-1), \forall i \in [k]$. If $R_L(f^*) = 0$, then, L is noise tolerant under class conditional noise when $\bar{\eta}_{ij} < (1 - \eta_i), \forall j \neq i$, $\forall i, j \in [k]$.*

Proof 3 *For class-conditional noise, we have*

$$\begin{aligned} R_L^\eta(f) &= \mathbb{E}_\mathcal{D}(1-\eta_{y_\mathbf{x}})L(f(\mathbf{x}), y_\mathbf{x}) + \mathbb{E}_\mathcal{D}\sum_{i\neq y_\mathbf{x}} \bar{\eta}_{y_\mathbf{x} i} L(f(\mathbf{x}), i) \\ &= \mathbb{E}_\mathcal{D}(1-\eta_{y_\mathbf{x}})(C - \sum_{i\neq y_x} L(f(\mathbf{x}), i)) \\ &\quad + \mathbb{E}_\mathcal{D}\sum_{i\neq y_x} \bar{\eta}_{y_\mathbf{x} i} L(f(x), i) \\ &= C\mathbb{E}_\mathcal{D}(1-\eta_{y_\mathbf{x}}) - \mathbb{E}_\mathcal{D}\sum_{i\neq y_\mathbf{x}} (1 - \eta_{y_\mathbf{x}} - \bar{\eta}_{y_\mathbf{x} i}) L(f(\mathbf{x}), i) \end{aligned} \tag{6}$$

Since f^*_η is the minimizer of R^η_L, we have $R^\eta_L(f^*_\eta) - R^\eta_L(f^*) \leq 0$ and hence from Eq.(6) we have

$$\mathbb{E}_\mathcal{D}\sum_{i\neq y_\mathbf{x}} (1 - \eta_{y_\mathbf{x}} - \bar{\eta}_{y_\mathbf{x} i})(L(f^*(\mathbf{x}), i) - L(f^*_\eta(\mathbf{x}), i)) \leq 0 \tag{7}$$

Since we are given $R_L(f^*) = 0$, we have $L(f^*(\mathbf{x}), y_\mathbf{x}) = 0$. Given the condition on L in the theorem, this implies $L(f^*(\mathbf{x}), i) = C/(k-1),\ i \neq y_\mathbf{x}$. As per the assumption on noise in the theorem, $(1 - \eta_{y_\mathbf{x}} - \bar{\eta}_{y_\mathbf{x} i}) > 0$. Also, L has to satisfy $L(f^*_\eta(\mathbf{x}), i) \leq C/(k-1),\ \forall i$. Thus for Eq.(7) to hold, it must be the case that $L(f^*_\eta(\mathbf{x}), i) = C/(k-1),\ \forall i \neq y_\mathbf{x}$ which, by symmetry of L, implies $L(f^*_\eta(\mathbf{x}), y_\mathbf{x}) = 0$. Thus minimizer of true risk is also a minimizer of risk under noisy data. This completes the proof.

Remark 3 Note that $1 - \eta_{y_\mathbf{x}} > \bar{\eta}_{y_\mathbf{x} i}$ implies $\eta_{y_\mathbf{x}} < (k-1)/k$. Thus the condition on noise rates for this theorem is more strict. For $i \neq j$, $\bar{\eta}_{ij}$ is the probability that a feature vector of class-i is labelled as class-j. If we set $\bar{\eta}_{ii} = 1 - \eta_i$ which is the probability of a feature vector of class-i having correct label, then the condition is that the matrix $[\bar{\eta}_{ij}]$ of label noise probabilities should be diagonal dominant. The condition on the loss function in the theorem is satisfied by some of the symmetric losses such as 0–1 loss and MAE loss. The condition that $R_L(f^*) = 0$ is restrictive. However, experimentally, even though minimum risk might not be 0, symmetric losses show good robustness even under class-conditional noise.

Some Loss Functions for Neural Networks

We assume standard neural network architecture with softmax output layer. If input to network is $\mathbf{x}$, then input to softmax layer is $f(\mathbf{x})$. Softmax layer computes:

$$u_i = \frac{\exp(f(\mathbf{x})_i)}{\sum_{j=1}^k \exp(f(\mathbf{x})_j)}, \quad i \in [k]$$

where $f(\mathbf{x})_i$ represents i^{th} component of $f(\mathbf{x})$. We have $\sum_{i=1}^k u_i = 1$. We define $\mathbf{u} = [u_1, \cdots, u_k]$. The label for the training patterns is in 'one-of-K' representation. If the class of $\mathbf{x}$ is j then $y_\mathbf{x}$ is given as $\mathbf{e_j}$ where $e_{ji} = 1$ if $i = j$, otherwise 0. We can now define some popular loss functions namely, categorical cross entropy (CCE), Mean square error (MSE) and Mean absolute error (MAE) as below.

$$L(f(\mathbf{x}), \mathbf{e_j}) = \begin{cases} \sum_{i=1}^k e_{ji} \log \frac{1}{u_i} = \log \frac{1}{u_j} & \text{CCE} \\ ||\mathbf{e_j} - u||_1 = 2 - 2u_j & \text{MAE} \\ ||\mathbf{e_j} - u||_2^2 = ||u||_2^2 + 1 - 2u_j & \text{MSE} \end{cases}$$

For these loss functions, we have

$$\sum_{i=1}^k L(f(\mathbf{x}), \mathbf{e_i}) = \begin{cases} \sum_{i=1}^k \log \frac{1}{u_i} & \text{CCE} \\ \sum_{i=1}^k (2 - 2u_i) = 2k - 2 & \text{MAE} \\ k||u||_2^2 + k - 2 & \text{MSE} \end{cases}$$

Thus, among these, only MAE satisfies symmetry condition given by Eq.(2). While MSE does not satisfy Eq.(2), one can show, using $\frac{1}{k} \leq ||u||_2^2 \leq 1$, that $k - 1 \leq \sum_i^k L_{mse}(f(x), i) \leq 2k - 2$. This boundedness makes it more robust than an unbounded loss such as CCE.

Informally, a loss function is said to be classification calibrated if a classifier having low enough risk under that loss would also have low risk under 0–1 loss. Logistic loss and exponential loss in multi-class settings have been proved to be classification calibrated (Weston and Watkins 1998; Tewari and Bartlett 2007; Bartlett, Jordan, and McAuliffe 2006). One can show that, MAE, MSE losses are also classification calibrated.

Consistency under Symmetric Label Noise

We showed that risk minimization with symmetric losses is robust to label noise. Since there is only a finite training set, one can only minimize Empirical Risk. We now prove consistency of empirical risk minimization under label noise.

Theorem 4 *Consider empirical risk minimization (ERM) under symmetric label noise over a given function class of finite VC dimension. If the loss L used for ERM is robust to label noise (i.e., satisfies eq. 2), then the error rate of minimizer of empirical risk with noisy samples converges uniformly to the error rate of the minimizer of risk under noise-free distribution.*

Proof 4

We denote by $er_\mathcal{D}[g]$ the error rate (i.e., 0-1 risk) of classifier g in the noise-free case. Under noise we denote it as $er_{\mathcal{D}_\eta}[g]$. Let $\hat{g}^*$ ($\hat{g}^*_\eta$) be the minimizer of empirical risk over n noise-free (noisy) samples. Let g^* (g^*_η) be the minimizer of risk. Since VC bounds are distribution independent, we have,

$$\begin{aligned} er_\mathcal{D}[\hat{g}^*] &\leq er_\mathcal{D}(g^*) + \epsilon(n, vc) \\ er_{\mathcal{D}_\eta}[\hat{g}^*_\eta] &\leq er_{\mathcal{D}_\eta}(g^*_\eta) + \epsilon(n, vc) \end{aligned}$$

Table 1: Standard Datasets and Architecture

Dataset (n_{tr}, n_{te}, c, d)	Hidden layer Architecture
MNIST $(60k, 10k, 10, 28 \times 28)$	Conv layer + max pooling $(dr = 0.25)$+ two layers 1024 units (dr=0.25, 0.5)
CIFAR 10 $(50k, 10k, 10, 3 \times 32 \times 32)$ (Krizhevsky and Geoffrey 2009)	2 Conv layers + max pooling (dr=0.2)+ 2 Conv layers + max-pooling (dr=0.2)+ 1 layer 512 units (dr=0.5)
Reuters RCV1 $(213k, 213k, 50, 2000)$(Lewis et al. 2004)	One layer 256 units (dr=0.5)
Reuters newswire $(8982, 2246, 46, 2k)$	One layer 128 units (dr=0.5)
20 newsgroup by-date $(11314, 7532, 20, 5k)$	Input directly connected to Softmax layer with max-norm constraint
Imdb Sentiment $(20k, 5k, 2, 5k)$ (Maas et al. 2011)	One embedding layer 50 units (dr=0.2)+ Conv layer + one layer 250 units (dr=0.5)

The term $\epsilon(n, vc)$ or simply ϵ goes to 0 with $\frac{vc}{n}$, where vc is the VC dimension of the function class (Vapnik 1995).

Under symmetric label noise η, we derived how error rate changes. (Note that 0-1 loss is symmetric). Thus,

$$er_{\mathcal{D}_\eta} = er_{\mathcal{D}}(1 - \frac{k\eta}{k-1}) + c'\eta$$

Then we have the following,

$$er_{\mathcal{D}_\eta}[\hat{g}_\eta^*] - er_{\mathcal{D}_\eta}[g_\eta^*] = (er_{\mathcal{D}}[\hat{g}_\eta^*] - er_{\mathcal{D}}[g_\eta^*])(1 - \frac{k\eta}{k-1}) \leq \epsilon$$

$$\Rightarrow \; er_{\mathcal{D}}[\hat{g}_\eta^*] - er_{\mathcal{D}}[g_\eta^*] = er_{\mathcal{D}}[\hat{g}_\eta^*] - er_{\mathcal{D}}[g^*] \leq \frac{\epsilon}{1 - \frac{k\eta}{k-1}}$$

where we have used $er_{\mathcal{D}}[g_\eta^*] = er_{\mathcal{D}}[g^*]$ which follows because L is robust to label noise. This completes the proof.

Empirical Results

In this section we illustrate the robustness of symmetric loss functions. We present results with two image data sets and four text data sets. In each case we learn a neural network classifier using the CCE, MSE and MAE loss functions. We add symmetric or class conditional noise with different noise rates to the training set. For learning, we minimize the empirical risk, with different loss functions, using stochastic gradient descent through backpropagation (Bergstra et al. 2010; Chollet 2015). The learnt networks are tested on noise-free test sets.

Experimental Setting

The specific image and text data sets used are shown in Table 1. In the table, for each data set, we mention size of training and test sets (n_{tr}, n_{te}), number of classes (c) and input dimension (d). Since some are image data while others are text data and feature space dimensions are all different, we have used different network architectures for each data set. These are also specified in Table 1. All networks used Rectified Linear Unit (ReLU) in the hidden layers and have softmax layer at the output with the size of the layer being the number of classes. All networks are trained through backpropagation with momentum term and weight decay. We have also used dropout regularization and the dropout rates are also shown in Table 1.

The results reported are averages over six runs. Label noise is added in the training set by changing the label of each example independently. For symmetric noise, we fix η and randomly change the label of each example. For class conditional noise, for each experiment, we generate a fixed label noise probability matrix, $[\bar{\eta}_{ij}]$, randomly and use that to decide the new labels. We ensure that the matrix is diagonal dominant as needed for our theoretical results.

Results and Discussion

In figure 1, we compare the robustness of MAE and CCE losses on MNIST image data set and RCV1 text data set. We have used symmetric label noise with $\eta = 0,\ 0.4,\ 0.8$. The figure shows the evolution of training and test accuracies of the network with number of epochs of training. As the graphs in Fig. 1(a)–(c) show, MAE loss is highly robust to symmetric label noise. The test accuracy achieved with MAE even under 80% noise is close to that with zero noise. On noise-free data, the accuracy achieved with CCE loss is a little bit higher than that with MAE. However, even at 40% noise, the test accuracy with CCE loss drops sharply. Similar trend can be seen on the RCV1 data. (See Fig. 1(e)–(g)). Here, eventhough the drop in accuracy of CCE loss is not as sharp, it is clearly seen that MAE is much more robust. In Fig. 1 (d) and (h) we show results under class-conditional noise (CC). In these problems we have no idea whether the minimum risk is zero. However, as can be seen from the figure, the symmetric MAE loss exhibits a good level of robustness under class conditional noise also.

Table 2, shows average test accuracy and standard deviation (over six runs) of the learnt networks for different noise rates. We show results for noise rate $\eta = 0.0,\ 0.3,\ 0.6$. (For the 2-class Imdb dataset, noise rate used is $\eta = 0.0,\ 0.15,\ 0.3$). We also show results for class conditional noise. As can be seen from the table, MAE exhibits good robustness. When the accuracy of MAE at 0% noise is high (e.g., MNIST and RCV1), it drops very little even under 60% noise rate. However, when accuracy achieved at 0% noise is poor (showing perhaps that the optimal risk is large or that number of examples is inadequate), the accuracy drops with noise. However, in all cases the drop in accu-

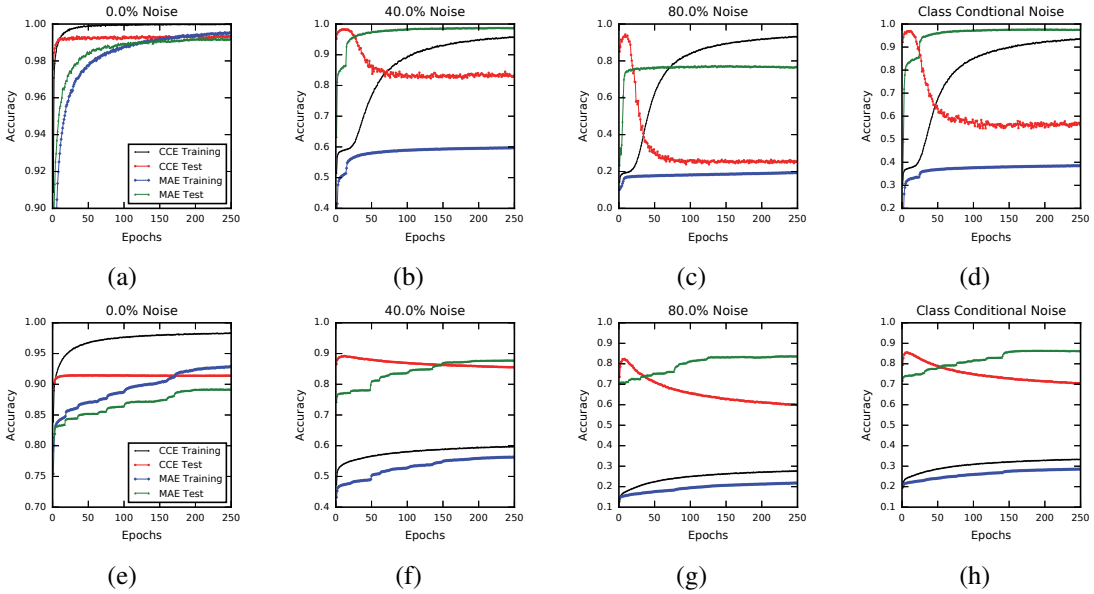

Figure 1: Train-Test Accuracies for log loss and MAE over epochs, for MNIST Datasets under (a) 0% noise (b) 40% noise (c) 80% noise (d) CC noise; and RCV1 Datasets under (e) 0% noise (f) 40% noise (g) 80% noise (h) CC noise. Legends shown in (a) and (e).

racy with MAE is much less than that with CCE. This is in accordance with our results on robustness of symmetric losses. We also see that robustness achieved by MSE (which is a bounded loss though it is not symmetric) is in between that of MAE and CCE. As can be seen from the last column of the table, MAE exhibits fair amount of robustness under class conditional noise also. We observed that higher dropout rates reduce sensitivity of CCE to label noise. This is expected as dropout works like regularizer (Srivastava et al. 2014). Interestingly, even without dropout, in many datasets MAE showed good robustness under label noise. On MNIST data with zero dropout, MAE retained almost same accuracy. We did not include these results in the table because it is customary to have high dropout rate.

Conclusion

In this paper, we derived some theoretical results on robustness of loss functions in multi-class classification. Such robust loss functions are useful because we can learn a good classifier (without any change in the algorithm or network architecture) even when training set labels are noisy. While we discussed these in the context of learning neural networks, our theoretical results are general and apply to any multi-class classifier learning through risk minimization. For learning neural networks, we showed that the commonly used CCE loss is sensitive to label noise while MAE loss is robust. We presented extensive empirical results to illustrate this. However, training a network under MAE loss would be slow because the gradient can quickly saturate while training. On the other hand, training under CCE is fast. Thus, designing better optimization methods for MAE is an interesting problem for future work. This would allow one to really exploit the robustness properties of MAE (and other such symmetric losses) proved here.

References

Angelova, A.; Abu-Mostafa, Y.; and Perona, P. 2005. Pruning training sets for learning of object categories. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 494–501.

Bartlett, P. L.; Jordan, M. I.; and McAuliffe, J. D. 2006. Convexity, classification, and risk bounds. *Journal of the American Statistical Association* 101(473):138–156.

Battista Biggio, B. N., and Laskov, P. 2011. Support vector machines under adversarial label noise. In *Proceedings of the Third Asian Conference on Machine Learning*, 97–112.

Bergstra, J.; Breuleux, O.; Bastien, F.; Lamblin, P.; Pascanu, R.; Desjardins, G.; Turian, J.; Warde-Farley, D.; and Bengio, Y. 2010. Theano: A cpu and gpu math compiler in python. In *Proc. 9th Python in Science Conf*, 1–7.

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

Brodley, C. E., and Friedl, M. A. 1999. Identifying mislabeled training data. *Journal Of Artificial Intelligence Research* 11:131–167.

Chollet, F. 2015. Keras. *GitHub repository: https://github.com/fchollet/keras*.

du Plessis, M. C.; Niu, G.; and Sugiyama, M. 2014. Analysis of learning from positive and unlabeled data. In *Advances in Neural Information Processing Systems*, 703–711.

Table 2: Accuracies under different noise rates (η) for all datasets (for Imdb, η's are halved). The last column gives accuracies under class conditional noise. In all the cases, standard deviation is shown only when it is more than 0.01

Data	loss	$\eta = 0\%$	$\eta = 30\%$	$\eta = 60\%$	CC
MNIST	CCE	0.9936	0.9138	0.5888	0.5775 ($\pm$0.0291)
	MAE	0.9916	0.9886	0.9799	0.9713
	MSE	0.9921	0.9868	0.9766	0.8505 ($\pm$0.0473)
RCV1	CCE	0.9126	0.8738	0.7905	0.7418 ($\pm$0.025)
	MAE	0.8732 ($\pm$0.0107)	0.8688	0.8637 ($\pm$0.0201)	0.8587
	MSE	0.9014	0.8943	0.8682 ($\pm$0.0120)	0.8315
Cifar 10	CCE	0.7812	0.5598 ($\pm$0.0170)	0.3083	0.4896
	MAE	0.7810 ($\pm$0.0190)	0.7011 ($\pm$0.0264)	0.5328 ($\pm$0.0251)	0.61425 ($\pm$0.0320)
	MSE	0.8074	0.7027	0.5257 ($\pm$0.0146)	0.6249 ($\pm$0.0359)
Imdb	CCE	0.8808	0.7729	0.6466	0.7858 ($\pm$0.0135)
	MAE	0.8813	0.8500	0.7352 ($\pm$0.0145)	0.8382 ($\pm$0.0127)
	MSE	0.8816	0.7725 ($\pm$0.0105)	0.6506 ($\pm$0.0103)	0.7874
News wire	CCE	0.7842	0.6905	0.4670	0.4973 ($\pm$0.0148)
	MAE	0.8081	0.7553	0.6357 ($\pm$0.0106)	0.6535
	MSE	0.7916	0.6626	0.4078 ($\pm$0.0172)	0.4377 ($\pm$0.0140)
News group	CCE	0.8006	0.7571	0.6435	0.5629
	MAE	0.7890	0.7749	0.7319	0.6772
	MSE	0.7999	0.7553	0.6347	0.5519

Frénay, B., and Verleysen, M. 2014. Classification in the Presence of Label Noise: A Survey. *IEEE Transactions on Neural Networks and Learning Systems* 25(5):845–869.

Ghosh, A.; Manwani, N.; and Sastry, P. 2015. Making risk minimization tolerant to label noise. *Neurocomputing* 160:93–107.

Jin, R.; Liu, Y.; Si, L.; Carbonell, J. G.; and Hauptmann, A. 2003. A new boosting algorithm using input-dependent regularizer. In *Proceedings of Twentieth International Conference on Machine Learning*.

Khardon, R., and Wachman, G. 2007. Noise tolerant variants of the perceptron algorithm. *Journal Of Machine Learning Research* 8:227–248.

Krizhevsky, A., and Geoffrey, H. 2009. Learning multiple layers of features from tiny images.

Lawrence, N. D., and Schölkopf, B. 2001. Estimating a kernel fisher discriminant in the presence of label noise. In *ICML*, volume 1, 306–313. Citeseer.

Lewis, D. D.; Yang, Y.; Rose, T. G.; and Li, F. 2004. Rcv1: A new benchmark collection for text categorization research. *Journal of machine learning research* 5(Apr):361–397.

Long, P. M., and Servedio, R. A. 2010. Random classification noise defeats all convex potential boosters. *Machine Learning* 78(3):287–304.

Maas, A. L.; Daly, R. E.; Pham, P. T.; Huang, D.; Ng, A. Y.; and Potts, C. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 142–150. Portland, Oregon, USA: Association for Computational Linguistics.

Manwani, N., and Sastry, P. 2013. Noise tolerance under risk minimization. *Cybernetics, IEEE Transactions on* 43(3):1146–1151.

Mnih, V., and Hinton, G. E. 2012. Learning to label aerial images from noisy data. In *ICML*. Citeseer.

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

Patrini, G.; Rozza, A.; Menon, A.; Nock, R.; and Qu, L. 2016. Making neural networks robust to label noise: a loss correction approach. *arXiv preprint arXiv:1609.03683*.

Reed, S.; Lee, H.; Anguelov, D.; Szegedy, C.; Erhan, D.; and Rabinovich, A. 2014. Training deep neural networks on noisy labels with bootstrapping. *arXiv preprint arXiv:1412.6596*.

Scott, C.; Blanchard, G.; and Handy, G. 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.

Srivastava, N.; Hinton, G. E.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research* 15(1):1929–1958.

Sukhbaatar, S.; Bruna, J.; Paluri, M.; Bourdev, L.; and Fergus, R. 2014. Training convolutional networks with noisy labels. *arXiv preprint arXiv:1406.2080*.

Tewari, A., and Bartlett, P. L. 2007. On the consistency of multiclass classification methods. *Journal of Machine Learning Research* 8(May):1007–1025.

van Rooyen, B.; Menon, A.; and Williamson, R. C. 2015. Learning with symmetric label noise: The importance of being unhinged. In *Advances in Neural Information Processing Systems*, 10–18.

Vapnik, V. N. 1995. *The Nature of Statistical Learning Theory*. New York, NY, USA: Springer-Verlag New York, Inc.

Weston, J., and Watkins, C. 1998. Multi-class support vector machines.

Xiao, T.; Xia, T.; Yang, Y.; Huang, C.; and Wang, X. 2015. Learning from massive noisy labeled data for image classification. In *The IEEE Conference on Computer Vision and Pattern Recognition*.

Zhu, X.; Wu, X.; and Chen, Q. 2003. Eliminating class noise in large datasets. In *Proceedings of the Twentieth International Conference on Machine Learning*, 920–927.