Robust Loss Functions under Multi-class Label Noise for Deep Neural Networks

Himanshu Kumar, P. S. Sastry
{himanshukr, sastry}@ee.iisc.ernet.in

Established 1911

Summary

- Robust learning of classifiers in presence of Label Noise.
- Under risk minimization (RM) framework, we prove sufficient conditions on a loss function for robust classifier learning.
- Theoretical results are illustrated with learning of Deep Neural Networks under label noise.
- A new loss is proposed for efficient learning of Deep Neural Networks under label noise.

Problem Definition

- We denote $\mathcal{X} \subset \mathbb{R}^d$ as the feature space of samples and $\mathcal{Y} = [k] = \{1, \cdots, k\}$ as class labels.
- $S = \{(\mathbf{x}_1, y_{\mathbf{x}_1}), \ldots, (\mathbf{x}_N, y_{\mathbf{x}_N})\} \in (\mathcal{X} \times \mathcal{Y})^N$ is a noise free training dataset, drawn *iid* according to an unknown distribution, $\mathcal{D}$, over $\mathcal{X} \times \mathcal{Y}$.
- A classifier: $h(\mathbf{x}) = \text{pred} \circ f(\mathbf{x})$ where $h : \mathcal{X} \to \mathcal{Y},\ f : \mathcal{X} \to \mathcal{C},\ \mathcal{C} \subseteq \mathbb{R}^k$.
(f itself is also referred to as the classifier.)
- The objective is to learn a classifier, f, which is a global minimizer of risk, R_L.

$$f^* = \arg\min_f R_L(f) = \arg\min_f \mathbb{E}_{\mathcal{D}}[L(f(\mathbf{x}), y_{\mathbf{x}})] \tag{1}$$

where $L : \mathcal{C} \times \mathcal{Y} \to \mathbb{R}^+$ is a loss function and $\mathbb{E}$ denotes expectation.

- $S_\eta = \{(\mathbf{x}_n, \hat{y}_{\mathbf{x}_n}), n = 1, \cdots, N\}$ is noisy training data available to the learner under noisy settings where,

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

and for all $\mathbf{x}$, conditioned on $y_{\mathbf{x}} = i$, with $\sum_{j \neq i} \bar{\eta}_{\mathbf{x}j} = \eta_{\mathbf{x}}$.

- L-risk of a classifier f under noisy setting is

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

where D_η is the joint distribution of $\mathbf{x}$, $\hat{y}_{\mathbf{x}}$. Let f_η^* be the global minimizer of $R_L^\eta(f)$.

- RM under loss function L, is said to be *noise-tolerant* if

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

Theorems

- **Def:** A loss function L is said to be *symmetric* if it satisfies, for some constant C,

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

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* $\eta < \frac{k-1}{k}$.

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* $\mathbf{x} \in \mathcal{X}$.

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* $\bar{\eta}_{ij} < (1 - \eta_i), \forall j \neq i, \forall i, j \in [k]$.

Results

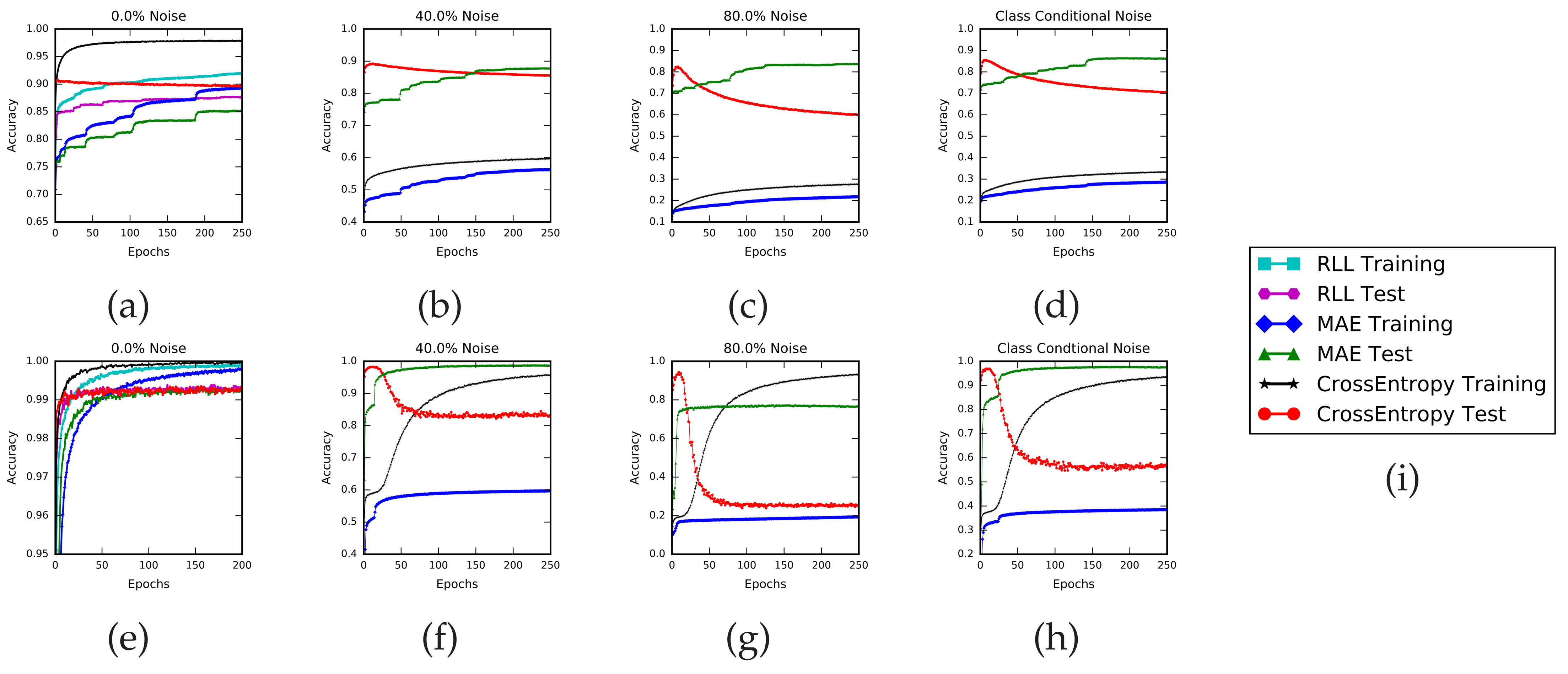

Figure 1: Train-Test Accuracies for CCE and MAE over epochs, for RCV1 Datasets under noise-rate (a) 0% (b) 40% (c) 80% (d) CC and MNIST Datasets under noise-rate (e) 0% (f) 40% (g) 80% (h) CC. Legends are shown in (i). (a), (e) also shows comparison between learning rate of RLL and MAE.

Types of Label Noise

- *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.
- *class-conditional* or asymmetric if $\eta_{\mathbf{x}} = \eta_{y_{\mathbf{x}}}$, and $\bar{\eta}_{\mathbf{x}j} = \bar{\eta}_{y_{\mathbf{x}},j}$.
- *non-uniform* if $\eta_{\mathbf{x}}$, $\bar{\eta}_{\mathbf{x}j}$ are functions of $\mathbf{x}$ and *simple non-uniform noise* if $\bar{\eta}_{\mathbf{x}j} = \frac{\eta_{\mathbf{x}}}{k-1}, \forall j \neq y_{\mathbf{x}}$.

Losses

- Categorical Cross Entropy (CCE)
- Mean Square Error (MSE)
- Mean Absolute Error (MAE)
- Robust Log Loss (RLL)

∗∗Deep Networks with softmax layer as the last layer, outputs a probability vector $\mathbf{u}$ for any $\mathbf{x}$ (i.e. $f(\mathbf{x}) = \mathbf{u}$). The final classifier would be $h(\mathbf{x}) = q$, where $q = \arg\max_{r \in [k]} u_r$

The loss functions are now defined in terms of $\mathbf{u}$ as

$$L(\mathbf{u}, \mathbf{e_j}) = \begin{cases} \sum_{i=1}^{k} e_{ji} \log \frac{1}{u_i} = \log \frac{1}{u_j} & \text{CCE} \\ ||\mathbf{e_j} - \mathbf{u}||_1 = 2 - 2u_j & \text{MAE} \\ ||\mathbf{e_j} - \mathbf{u}||_2^2 = ||\mathbf{u}||_2^2 + 1 - 2u_j & \text{MSE} \\ log2 - e_{jj} \log(1 + u_i) + & \\ \quad \sum_{\substack{i=1 \\ i \neq j}}^{k} \frac{e_{jj}}{k-1} \log(1 + u_i) & \text{RLL} \end{cases}$$

For these loss functions, we have

$$\sum_{i=1}^{k} L(\mathbf{u}, \mathbf{e_i}) = \begin{cases} \sum_{i=1}^{k} \log \frac{1}{u_i} & \text{CCE} \\ \sum_{i=1}^{k} (2 - 2u_i) = 2k - 2 & \text{MAE} \\ k||\mathbf{u}||_2^2 + k - 2 & \text{MSE} \\ klog2 & \text{RLL} \end{cases}$$

MAE, RLL satisfies our symmetry condition.

Results

Data	loss	$\eta = 0\%$	$\eta = 30\%$	$\eta = 60\%$	CC
MNIST	CCE	**0.9935**	0.8955	0.5845	0.5776
	MAE	0.9924	**0.9900**	**0.9788**	0.9313
	MSE	0.9921	0.9868	0.9766	0.8505
	RLL	0.9934	0.9896	0.9639	**0.9455**
RCV1	CCE	**0.9078**	0.7630	0.5321	0.4920
	MAE	0.8627	0.8431	**0.8401**	**0.8269**
	MSE	0.9014	**0.8743**	0.8382	0.8015
	RLL	0.8876	0.8592	0.8254	0.8141
Imdb	CCE	0.8645	0.72316	0.6268	0.7858
	MAE	0.8520	**0.8088**	**0.7174**	0.8282
	MSE	0.8616	0.7725	0.6506	0.7874
	RLL	**0.8648**	0.8020	0.7010	**0.8348**
News group	CCE	0.7913	0.6918	0.4953	0.3771
	MAE	**0.8048**	0.7742	0.6665	0.5547
	MSE	0.7999	0.7553	0.6347	0.5519
	RLL	0.7929	**0.7796**	**0.6990**	**0.6019**

Table 1: Accuracies under different noise rates (η) for all datasets (for Imdb, η's are halved). The last column gives accuracies under class conditional noise.

Conclusion

- RM with symmetric losses has interesting robustness properties.
- ERM is shown to be consistent under uniform noise.
- Results with MAE, RLL show robustness under uniform and class conditional noise (with diagonally dominant noise matrix).
- $R_L(f^*) = 0$ for robustness under non-uniform noise is very restrictive.
- Learning with RLL is faster compared to MAE but slower compared to CCE. Further work on optimization algorithms for fast learning is required.