Learning from Noisy Labels

Ajay S. Joshi(CS15B047)

IIT Madras

June 8, 2020

Contents

Introduction

Problem Statement: Developing algorithms to make machine learning models robust to incorrect labels in datasets.

Motivation:

- Large scale labeled datasets with precise annotations are required for learning ML models, but labeling is expensive, time-consuming and potentially erroneous
- Crowdsourcing can be used for labeling or internet data(eg: images) can be used by extracting labels from queries/tags, both of which will introduce noise into labels

Related Work

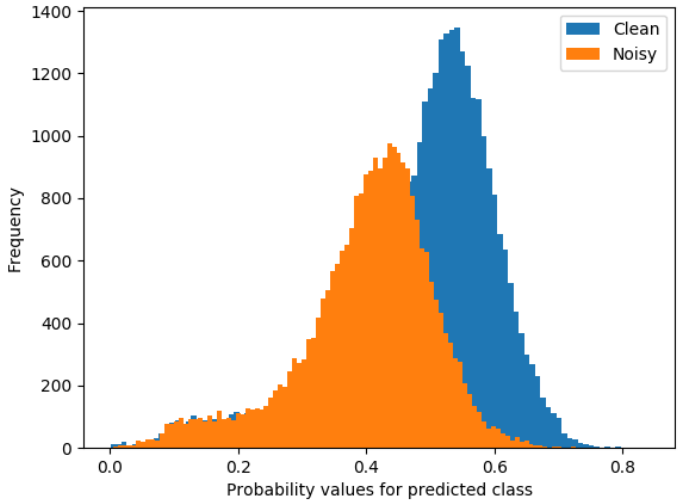
- Theoretical Approaches
- Noise Modelling
- Instance Selection

Contributions

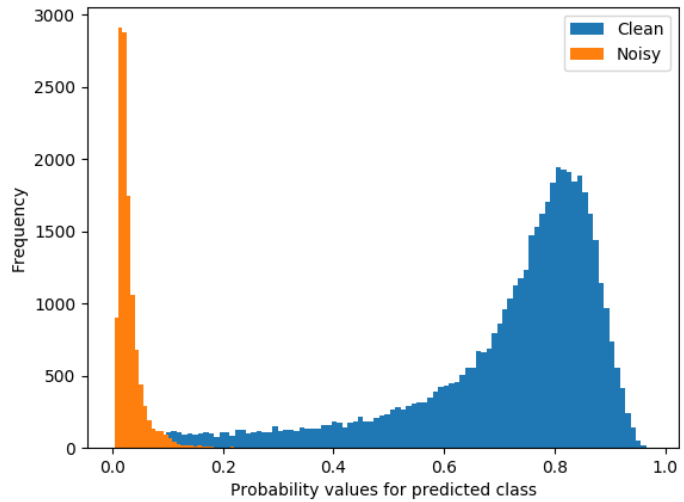
1. **Characteristics** differentiating clean and noisy examples
2. **Novel Observation** regarding variation of model performance with model complexity
3. Iterative Reweighting **Algorithm**

Characteristics of clean and noisy examples: I

- **Predicted probability** for class label can be used a characteristic for separating clean and noisy examples



(a) High Noise

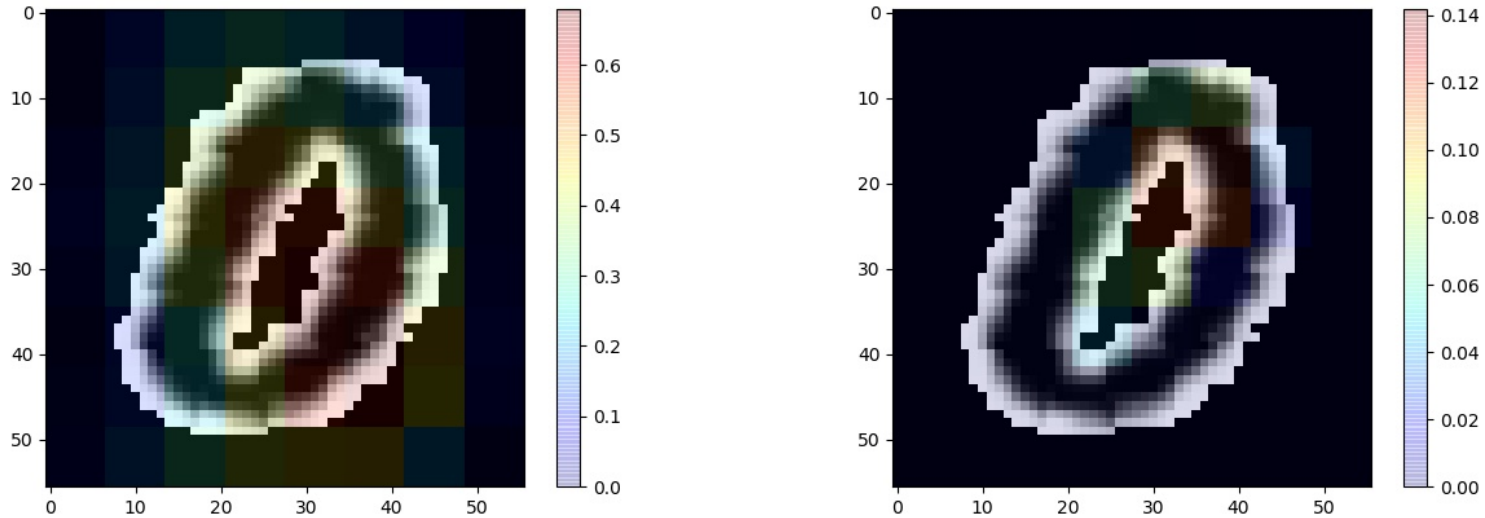


(b) Low Noise

Figure: Probability histograms after an epoch for the MNIST dataset.

Characteristics of clean and noisy examples: II

- Class Activation Maps for examples with correct labels are more focused on the important image regions, whereas they are random for examples with noisy labels

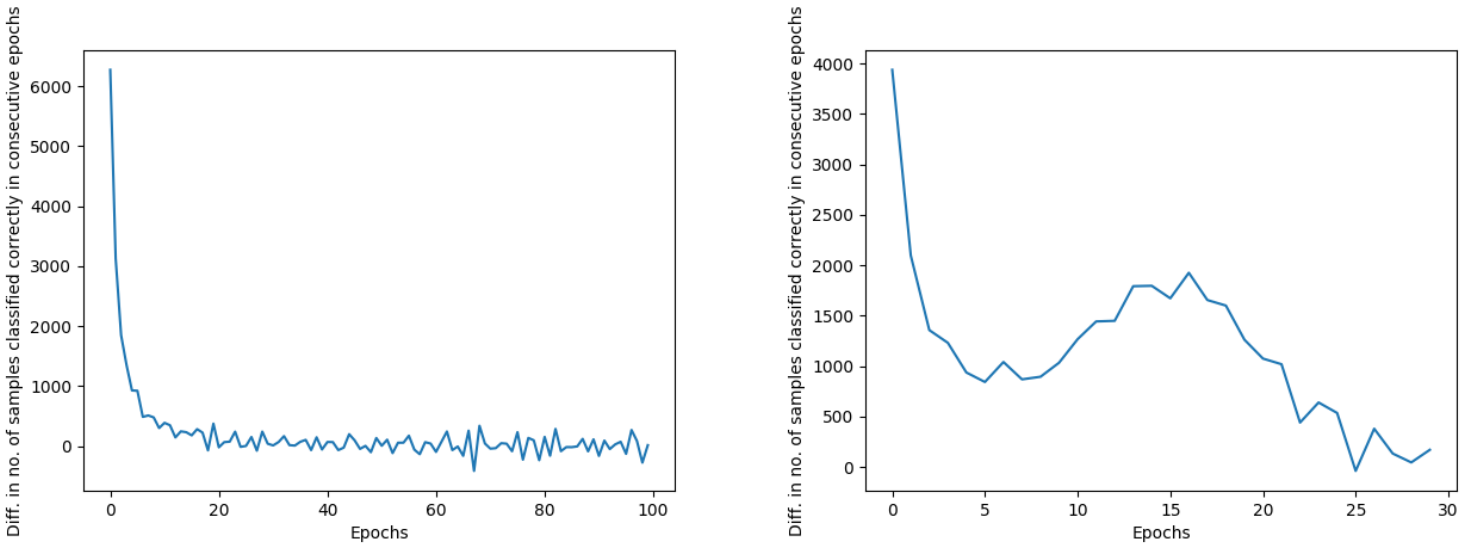


(a) CAM for class 0 (correct label) (b) CAM for class 1 (wrong label)

Figure: Class Activation Maps from the MNIST dataset.

Variation of Model Performance with Model Complexity: Training Steps

- Difference in model performance varies differently with varying model complexity with and without label noise

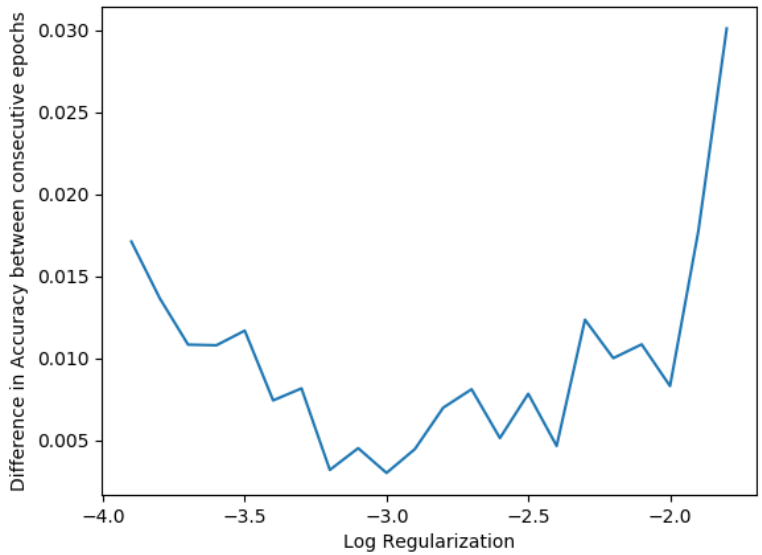


(a) Label noise absent

(b) Label noise present

Figure: Difference in examples classified between consecutive epochs for CIFAR-10 dataset.

Variation of Model Performance with Model Complexity: Regularization



(a) CIFAR 10- high noise

Figure: Plots show the same quantities as Figure 3, but use regularization to control model complexity

Iterative Reweighting Algorithm

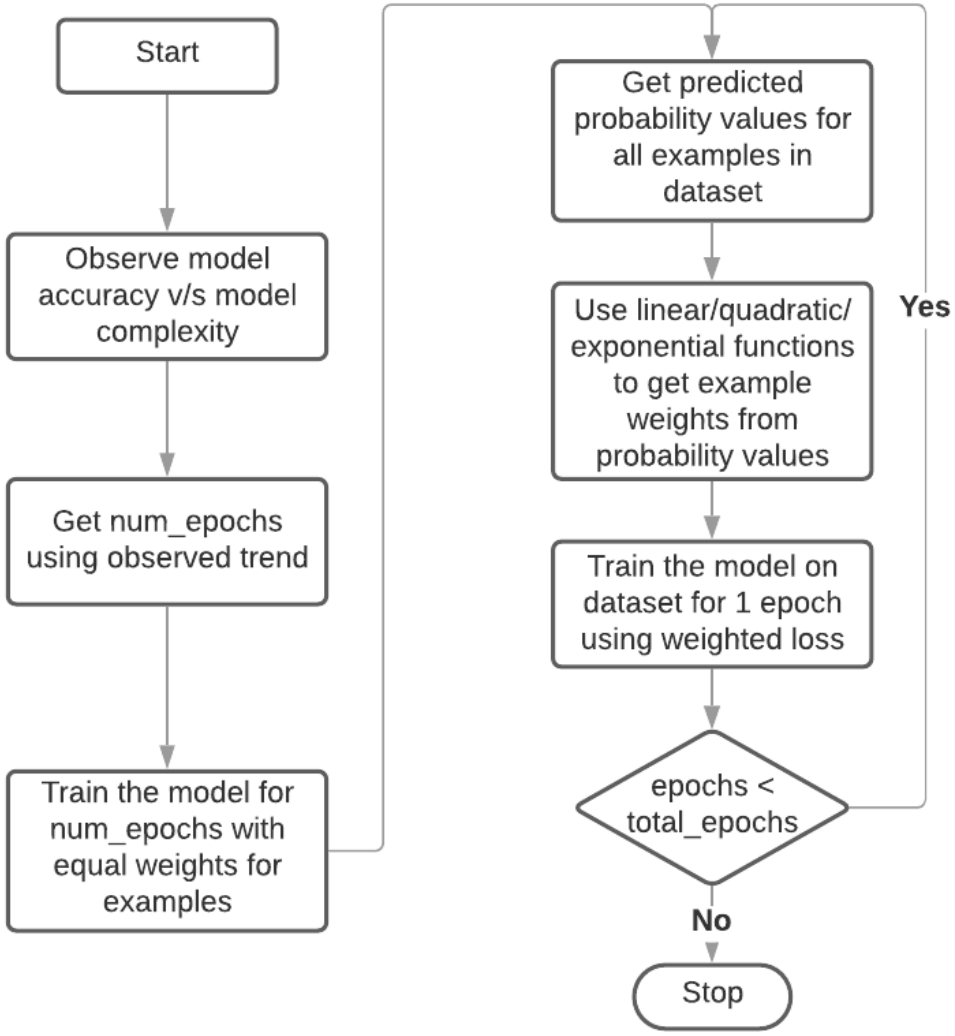


Figure: Flowchart for the iterative reweighting algorithm

Results: Test Accuracy

Method\ Noise	Pair-45	Symmetric-50	Symmetric-20
F-Correction [3]	0.24	79.61	98.80
Mentornet [2]	80.88	90.05	96.70
Co-teaching [1]	87.63	91.32	97.25
Ours (5)	**98.76**	**98.9**	**99.42**

Table: Comparison of Results on MNIST test data after 200 epochs

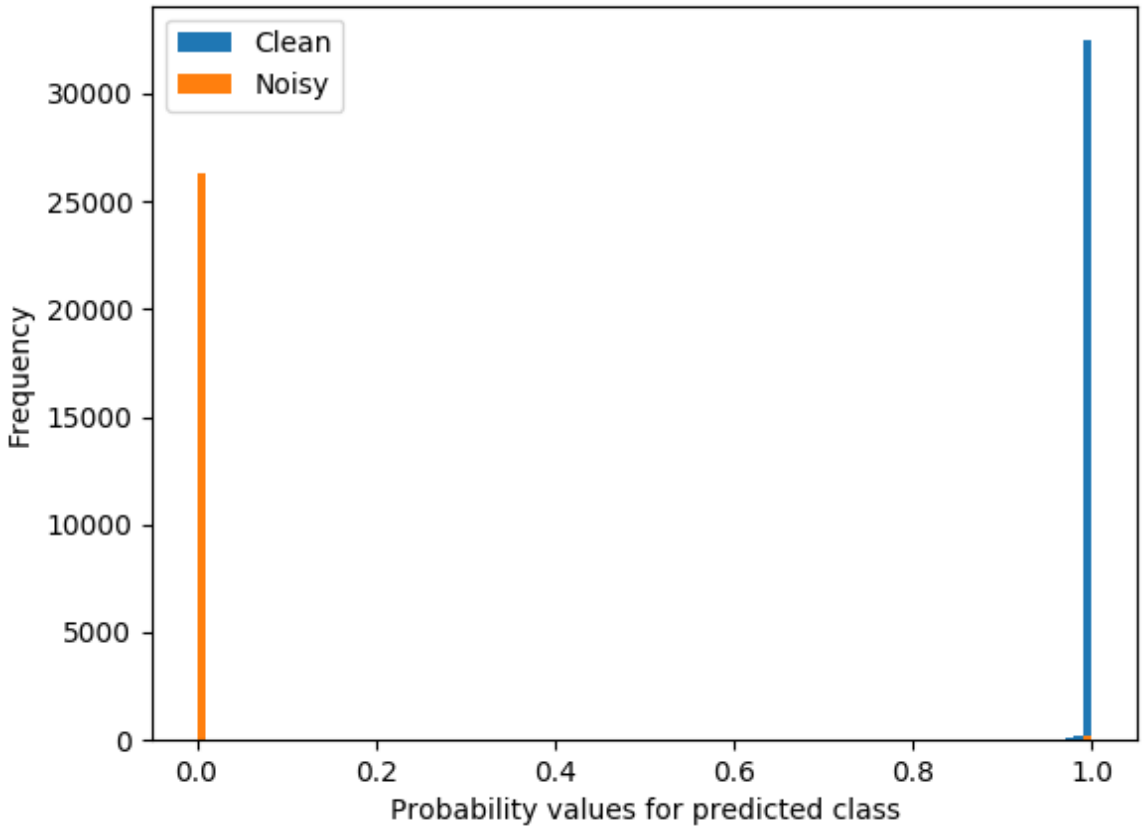
Method\ Noise	Pair-45	Symmetric-50	Symmetric-20
F-Correction [3]	6.61	59.83	84.55
Mentornet [2]	58.61	71.10	80.76
Co-teaching [1]	72.62	74.02	82.32
Ours (5)	**81.77**	**81.57**	**87.7**

Table: Comparison of Results on CIFAR-10 test data after 200 epochs

Method\ Noise	Pair-45	Symmetric-50	Symmetric-20
F-Correction [3]	1.60	41.04	**61.87**
Mentornet [2]	31.60	39.00	52.13
Co-teaching [1]	34.81	**41.37**	54.23
Ours (5)	**40.00**	39.77	58.07

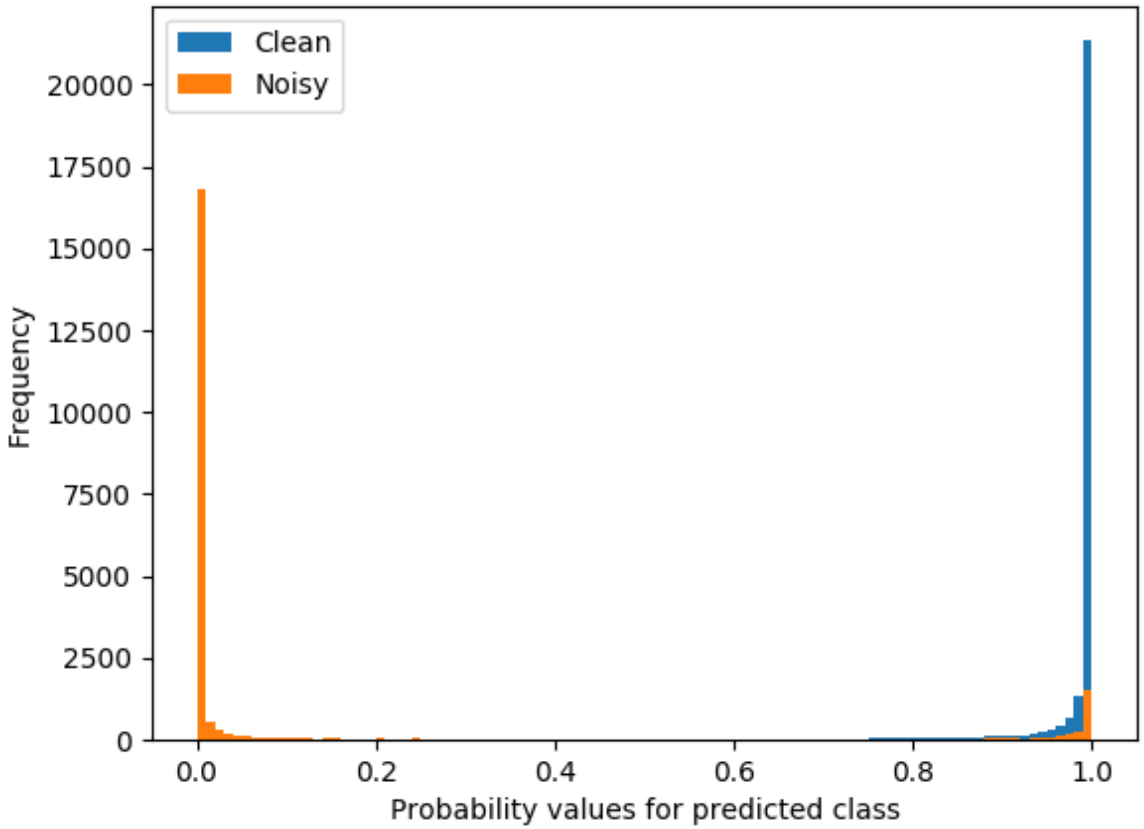
Table: Comparison of Results on CIFAR-100 test data after 200 epochs

Results: Probability Histograms I



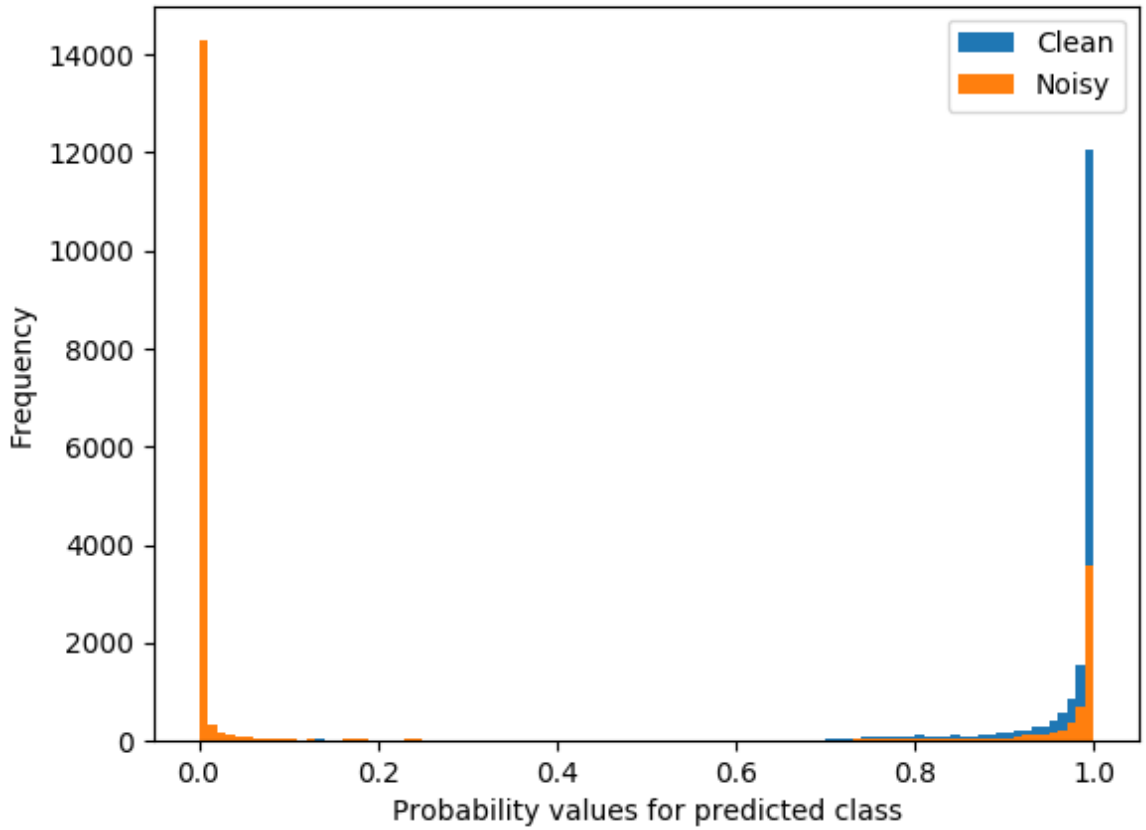
(a) MNIST Dataset

Results: Probability Histograms II



(b) CIFAR-10 Dataset

Results: Probability Histograms III



(c) CIFAR-100 Dataset

Figure: Probability histograms after training for high noise case

Generalization to Real-World Noise

- Currently, not aware of any previous work explaining generalization of model across noisy labeled datasets
- Experiments using recent methods like [1], [5] do not show encouraging performance on real-world noisy dataset
- Preliminary results of our algorithms show that we don't achieve any significant improvement over the baseline

Dataset\ Algorithm	Cross-Entropy	Co-Teaching [1]	Co-Teaching+ [5]	**Ours (5)**
Clothing-1M[4]	69.54%	68.11	41.32	**70.11**

Table: Results of papers [1], [5] on the Clothing-1M dataset, which is a real-world noisy dataset

Conclusion and Future Work

- Defined two distinguishing characteristics: class-prediction probability, quality of CAM visualizations
- Made novel observation regarding variation of model performance with model complexity
- Formulated Iterative Reweighting Algorithm which performs better than many recent methods by significant margins

Future Work:

- Generalization of models over data with different noise
- Using CAMs to develop algorithms for learning from noisy labels

References I


B. Han, Q. Yao, X. Yu, G. Niu, M. Xu, W. Hu, I. Tsang, and M. Sugiyama.
Co-teaching: Robust training of deep neural networks with extremely noisy labels.
In *Advances in neural information processing systems*, pages 8527–8537, 2018.


L. Jiang, Z. Zhou, T. Leung, L.-J. Li, and L. Fei-Fei.
Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels.
arXiv preprint arXiv:1712.05055, 2017.


G. Patrini, A. Rozza, A. Krishna Menon, R. Nock, and L. Qu.
Making deep neural networks robust to label noise: A loss correction approach.
In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1944–1952, 2017.

References II


T. Xiao, T. Xia, Y. Yang, C. Huang, and X. Wang.
Learning from massive noisy labeled data for image classification.
In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2691–2699, 2015.


X. Yu, B. Han, J. Yao, G. Niu, I. W. Tsang, and M. Sugiyama.
How does disagreement help generalization against label corruption?
arXiv preprint arXiv:1901.04215, 2019.

Thank You !

Algorithm based on Regularization

Input: Noisy data: D $= \{x_i, y_i\}_{i=1}^{i=N}$

Hyperparameters: Regularization values V in descending order, such that $|V| = k$, Epochs = E

Output: Example weights W, a vector of length N, with $0 \leq W_i \leq 1 \; \forall i \; 1 \leq i \leq N$

Initialize DNN model parameters θ

$T \leftarrow 0$ //Denotes number of examples which are given non-zero weight

$\forall x \in \{1, \cdots, N\}, W_x \leftarrow 0$

for $i \leftarrow 1$ **to** k **do**

  $\lambda \leftarrow V[i]$

  for $e \leftarrow 1$ **to** k **do**

    Train the model with parameters θ for 1 epoch using Cross-Entropy loss and L2 regularization with parameter λ.

    $X_{curr} \leftarrow$ Set of examples classified correctly at end of epoch e

    $X_{prev} \leftarrow$ Set of examples having non-zero weight at end of epoch $e - 1$

    $X_{new} \leftarrow X_{curr} \setminus X_{prev}$

    w $\leftarrow$ 1 - $\frac{T}{N}$

    $\forall x \in X_{new}, W_x \leftarrow w$

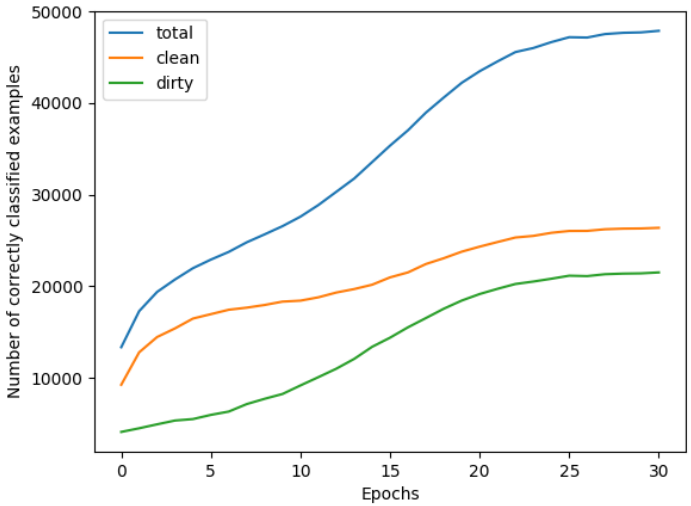
    $T \leftarrow T + |X_{new}|$

  end

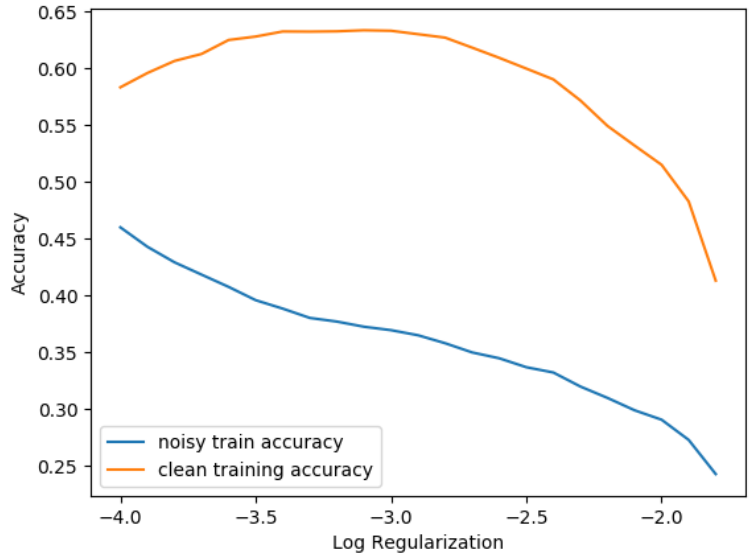
end

Algorithm 1: Weighting examples using Regularization

Plots regarding novel observation



(a) Evidence for Novel observation : training epochs



(b) Evidence for Novel observation : regularization

Figure: Evidence for the novel observation showing training accuracy with epochs