DISPARATE IMPACT WHEN LEARNING WITH NOISY LABELS

Yang Liu
yangliu@ucsc.edu
CSE, UC Santa Cruz

ABSTRACT

This paper aims to provide understandings for the effect of an over-parameterized model, e.g. a deep neural network, memorizing instance-dependent noisy labels. We first quantify the harms caused by memorizing noisy instances from different spectra of the sample distribution. We then analyze how several popular solutions for learning with noisy labels mitigate this harm at the instance-level. Our analysis reveals that existing approaches handle noisy instances differently. While higher-frequency instances often enjoy a high probability of an improvement by applying these approaches, lower-frequency instances do not. This observation requires us to rethink the distribution of label noise across instances and might potentially require different treatments for instances in different regimes.

1 INTRODUCTION

A salient feature of a over-parameterized model, e.g. a deep neural network, is its ability to memorize examples Zhang et al. (2016); Neyshabur et al. (2017), and the memorization has proven to benefit the generalization Arpit et al. (2017); Feldman (2020); Feldman & Zhang (2020). Nonetheless, the potential existence of label noise, combined with the memorization effect, might lead to detrimental consequence Song et al. (2020); Yao et al. (2020); Cheng et al. (2020b); Chen et al. (2019); Han et al. (2020); Song et al. (2020). In light of the reported empirical evidence of harms caused by over-memorizing noisy labels, we set out to understand this effect analytically. Built on a recent analytical framework Feldman (2020), we demonstrate the varying effects of memorizing noisy labels associated with instances that sit at the different spectra of the sample distribution. In particular, we prove that:

Theorem 1 (Informal)**.** *For a sample x with true label y that appears l times in the training data (with n samples), a deep neural network h memorizing its l noisy labels leads to the following order of excessive generalization error:*

$$\text{Excessive Generalization Error} = \Omega\left(\frac{l^2}{n^2} \cdot \textit{(noise rate at } x\textit{)}\right)$$

Soon since the above negative effect was empirically shown, learning with noisy labels has been recognized as a challenging and important task. The literature has observed growing interests in proposing defenses, see Natarajan et al. (2013); Liu & Tao (2016); Menon et al. (2015); Liu & Guo (2020); Lukasik et al. (2020) and many more. The second contribution of this paper is to build an analytical framework to gain new understandings of how the existing solutions fare. While most existing theoretical results focus on the setting where label noise is homogeneous across training examples and focus on the distribution-level analysis, ours invests on the instance-level and aims to quantify when these existing approaches work and when they fail for different regimes of instances. Our result points out a salient disparate effect that while noisy labels for highly frequent samples contribute more to the drop of generalization power, they are also the easier cases to fix with; on the other hand, we prove that existing solutions can have a substantial probability to fail at the long-tail examples Zhu et al. (2014):

Theorem 2 (Informal)**.** *Suppose a sample x appears l times in the training data. When l is large (high-frequency samples), with high probability, performing loss correction Natarajan et al. (2013); Patrini et al. (2017) and using peer loss correction Liu & Guo (2020) on x improves generalization*

error compared to memorizing the noisy labels. When l is small (low-frequency samples), with non-negligible probability (bounded below from 0), both loss correction and peer loss incur higher prediction error on x.

Our results call for a hybrid treatment of noisy instances. Due to space limit, this workshop manuscript focuses on presenting our main results for the disparate impact when learning with noisy labels. All relevant and missing details can be found in our full paper:
Liu (2021) [https://arxiv.org/abs/2102.05336].

2 FORMULATION

To reuse the main analytical framework built in Feldman (2020), we follow their notations. In the clean setting, a training dataset $S = \{(x_1, y_1), ..., (x_n, y_n)\}$ is available. Each x indicates a feature vector and each y is an associated label. Denote by X the space of x and Y the space for y. Jointly (x, y) are drawn from an unknown distribution $\mathcal{P}$ over $X \times Y$. Specifically, x is sampled from a distribution $\mathcal{D}$, and the true label y for x is specified by a function $f : X \to Y$ drawn from a distribution $\mathcal{F}$. The learner's algorithm $\mathcal{A}$, as a function of the training data S, returns a classifier or function $h : X \to Y$. We define the following generalization error $\mathsf{err}_{\mathcal{P}}(\mathcal{A}, S) := \mathbb{E}_{h\sim\mathcal{A}(S)}[\mathsf{err}_{\mathcal{P}}(h)]$, where $\mathsf{err}_{\mathcal{P}}(h) := \mathbb{E}_{\mathcal{P}}[\mathbb{1}(h(x) \neq y)]$. When there is no confusion, we shall use x, y to denote the random variables generating these quantities when used in a probability measure. To better and clearly demonstrate the main message of this paper, we consider discrete domains of X and Y such that $|X| = N, |Y| = m$. Our model, as well as the main generalization results, can mostly extend to a setting with continuous X (Section 4, Feldman (2020)). We briefly discuss it after we introduce the following process to capture the generation of each instance x: We follow Feldman (2020) to consider the following model to characterize an unstructured discrete domain of classification problems:

- Let $\pi = \{\pi_1, ..., \pi_N\}$ denote the priors for each $x \in X$.
- For each $x \in X$, sample a quantity p_x independently and uniformly from the set π.
- Then the resulting probability mass function of x is given by $D(x) = \frac{p_x}{\sum_{x\in X} p_x}$ - this forms the distribution $\mathcal{D}$ that x will be drawn from.

For the case with continuous X, instead of assuming a prior π over each x in a finite X, it's assumed there is a prior π defined over N mixture models. Each x has a certain probability of being drawn from each model and then will realize according to the generative model. Each of the generative models captures similar but non-identical examples. With the above generation process, denote by $\mathbb{P}[\cdot|S]$ the marginal distribution over $\mathcal{P}$ conditional on S, we further define the following conditional generalization error (on the realization of the training data S):

$$\mathsf{err}(\pi, \mathcal{F}, \mathcal{A}|S) := \mathbb{E}_{\mathcal{P}\sim\mathbb{P}[\cdot|S]}\left[\mathsf{err}_{\mathcal{P}}(\mathcal{A}, S)\right].$$

***l*-appearance samples**: We denote by $X_{S=l}$ as the set of xs that appeared exactly l times in the dataset S. The difference in l helps us capture the imbalance of the distribution of instances. Later we show that the handling of samples with different frequencies matters differently.

2.1 NOISY LABELS

We consider a setting where the training labels are noisy. Suppose for each training instance (x, y), instead of observing the true label y, we observe a noisy copy of it, denoting by $\tilde{y}$. Each $\tilde{y}$ is generated according to the following model:

$$T_{k,k'}(x) := \mathbb{P}[\tilde{y} = k'|y = k, x], k', k \in Y. \tag{1}$$

We will denote by $T(x) \in \mathbb{R}^{m\times m}$ the noise transition matrix with the (k, k')-entry defined by $T_{k,k'}(x)$. Each of the above noisy label generation is independent across different x. We will denote the above noisy dataset as $\tilde{S} := \{(x_1, \tilde{y}_1), ..., (x_n, \tilde{y}_n)\}$. For an x that appears l times in the dataset, it will have l independently generated noisy labels. One can think of these as l similar data instances, with each of them equipped with a single noisy label collected independently. $T(x)$ varies across the dataset S, and possibly that $T(x)$ would even be higher for low-frequency/rare samples, due to its inherent difficulties in recognizing and labeling them.

3 IMPACT OF MEMORIZING NOISY LABELS

In this section, we discuss the impacts of noisy labels when training a model that can memorize examples. Our analysis builds on a recent generalization bound for studying the memorization effects of a over-parameterized model.

3.1 GENERALIZATION

Denote by the $\bar{\pi}^N$ as the resulting marginal distribution over x: $\bar{\pi}^N(\alpha) := \mathbb{P}[D(x) = \alpha]$. $\bar{\pi}^N$ controls the true frequency of generating samples. Define the following quantity:

$$\tau_l := \frac{\mathbb{E}_{\alpha\sim\bar{\pi}^N}[\alpha^{l+1}\cdot(1-\alpha)^{n-l}]}{\mathbb{E}_{\alpha\sim\bar{\pi}^N}[\alpha^{l}\cdot(1-\alpha)^{n-l}]} \tag{2}$$

Theorem 2.3 of Feldman (2020) provides the following generalization error of an algorithm $\mathcal{A}$:

Theorem 3 (Theorem 2.3, Feldman (2020))**.** *For every learning algorithm $\mathcal{A}$ and every dataset $S \in (X \times Y)^n$:*

$$\mathsf{err}(\pi, \mathcal{F}, \mathcal{A}|S) \geq opt(\pi, \mathcal{F}|S) + \sum_{l\in[n]} \tau_l \cdot \sum_{x\in X_{S=l}} \mathbb{P}_{h\sim\mathcal{A}}[h(x) \neq y], \tag{3}$$

where in above, $opt(\pi, \mathcal{F}|S) := \min_{\mathcal{A}} \mathsf{err}(\pi, \mathcal{F}, \mathcal{A}|S)$, the minimum achievable generalization error.

We will focus on building our results and discussions using this generalization bound. In particular, our discussion will focus on how label noise can disrupt the training of a model through the changes of the following **Excessive Generalization Error:**

$$\mathsf{err}^+(\mathcal{P}, \mathcal{A}|S) := \sum_{l\in[n]} \tau_l \sum_{x\in X_{S=l}} \mathbb{P}_{h\sim\mathcal{A}(S')}[h(x) \neq y], \tag{4}$$

We will also denote by $\mathsf{err}_l^+(\mathcal{P}, \mathcal{A}, x|S) := \tau_l \cdot \mathbb{P}_{h\sim\mathcal{A}(S')}[h(x) \neq y]$, the **Individual Excessive Generalization Error** caused by a $x \in X_{S=l}$. Easy to see that $\mathsf{err}^+(\mathcal{P}, \mathcal{A}|S) = \sum_x \mathsf{err}_l^+(\mathcal{P}, \mathcal{A}, x|S)$.

3.2 IMPORTANCE OF MEMORIZING A l-APPEARANCE SAMPLE

Define $\mathsf{weight}\big(\pi, [\beta_1, \beta_2]\big)$ the expected fraction of distribution D contributed by frequencies in the range $[\beta_1, \beta_2]$: $\mathsf{weight}\big(\pi, [\beta_1, \beta_2]\big) := \mathbb{E}\left[\sum_{x\in X} D(x)\cdot 1(D(x) \in [\beta_1, \beta_2])\right]$. We bound τ_l for an arbitrary l:

Theorem 4. *For sufficiently large n, N, when $\pi_{max} \leq \frac{1}{20}$:*

$$\tau_l \geq 0.4\frac{l(l-1)}{n(n-1)} \cdot \mathit{weight}\left(\pi, \left[\frac{2}{3}\frac{l-1}{n-1}, \frac{4}{3}\frac{l}{n}\right]\right) \tag{5}$$

We observe that $\frac{l(l-1)}{n(n-1)} = O(\frac{l^2}{n^2})$.

3.3 IMPACT OF MEMORIZING NOISY LABELS

To study the negative effects of memorizing noisy labels, we first define memorization of noisy labels. For a $x \in X_{S=l}$ and its associated l noisy labels, denote by $\tilde{\mathbb{P}}[\tilde{y} = k|x], k \in Y$ the empirical distribution of the l noisy labels: for instance when $l = 3$, and two noisy labels are 1, then $\tilde{\mathbb{P}}[\tilde{y} = 1|x] = \frac{2}{3}$.

Definition 1 (Memorization of noisy labels)**.** *We call a model h memorizing noisy labels for instance x if $\mathbb{P}_{h\sim\mathcal{A}(S')}[h(x) = k] = \tilde{\mathbb{P}}[\tilde{y} = k|x]$.*

Effectively the assumption states that when, e.g. a deep neural network memorizes all l noisy labels for instance x, its output will follow the same empirical distribution. It has been shown in the literature Cheng et al. (2020b;a) that a fully memorizing neural network will be able to encode $\tilde{\mathbb{P}}[\tilde{y} = k|x]$ for each x. Then based on Theorem 4, we summarize our first observation that over-memorizing noisy labels for higher frequency samples will lead to bigger drop in the generalization power:

Theorem 5. *For $x \in X_{S=l}$ with true label y, h memorizing its l noisy labels leads to the following order of individual excessive generalization error* $\mathsf{err}_l^+(\mathcal{P}, \mathcal{A}, x|S)$*:*

$$\Omega\left(\frac{l^2}{n^2}\textit{weight}\left(\pi, \left[\frac{2}{3}\frac{l-1}{n-1}, \frac{4}{3}\frac{l}{n}\right]\right) \cdot \sum_{k\neq y}\tilde{\mathbb{P}}[\tilde{y}=k|x]\right)$$

4 Disparate Effect when Learning with Noisy Labels

In this section, we revisit 1) loss correction Natarajan et al. (2013); Patrini et al. (2017), a popular approach, and 2) peer loss Liu & Guo (2020), a recently proposed approach that does not rely the knowledge of transition matrix, and study how they offer fixes and under what conditions they might fail to work. For readers who are interested in the details of the two approaches, please also refer to our full version of this workshop manuscript Liu (2021).

Unless stated otherwise, throughout the section, we focus on a particular instance $x \in X_{S=l}$ with true label y and l corresponding noisy labels $\tilde{y}$s, one for each appearance. For a more clear demonstration, let's focus on the binary case that $y \in \{-1, +1\}$. Consider a particular x, and $T(x)$:

$$T(x) := \begin{bmatrix} 1-e_- & e_- := \mathbb{P}[\tilde{y}=+1|y=-1] \\ e_+ & 1-e_+ := \mathbb{P}[\tilde{y}=+1|y=+1] \end{bmatrix} \tag{6}$$

4.1 Loss correction

The first message we are ready to send is: **Loss correction returns better generalization power than memorizing noisy labels directly with high probability for instances with large l [high-frequency instances]**.

Theorem 6. *W.p. at least $1-e^{-2l(1/2-e_{sgn(y)})^2}$, performing loss correction for $x \in X_{S=l}$ improves the excessive generalization error* $\mathsf{err}_l^+(\mathcal{P}, \mathcal{A}, x|S)$ *by*

$$\Omega\left(\frac{l^2}{n^2} \cdot \textit{weight}\left(\pi, \left[\frac{2}{3}\frac{l-1}{n-1}, \frac{4}{3}\frac{l}{n}\right]\right)\right)$$

Our next message is: **Loss correction fails with substantial probability for x with small l [low-frequency instances]**. Denote by $D_{\mathrm{KL}}(\frac{1}{2}||e)$ the Kullback-Leibler distance between two Bernoulli 0/1 random variables of parameter $1/2$ and e, we prove:

Theorem 7. *For $x \in X_{S=l}$, w.p. at least $\frac{1}{\sqrt{2l}} \cdot e^{-l \cdot D_{\mathrm{KL}}\left(\frac{1}{2}||e_{sgn(y)}\right)}$, h memorizing $\mathbf{y}_{\mathrm{LC}}$ returns larger $\mathbb{P}[h(x) \neq y]$ than memorizing the noisy label $\tilde{\mathbf{y}}$.*

4.2 Peer loss

The first message we send for peer loss is that **peer loss extremizes h's prediction to the correct label with high probability for x with large l**. Consider the binary classification case, and an example $x \in X_{S=l}$ with true label y. Denote by $p_+ := \mathbb{P}_{y' \in \mathcal{F}|S}[y'=+1], p_- := \mathbb{P}_{y' \in \mathcal{F}|S}[y'=-1]$.

Theorem 8. *W.p. at least $1-e^{\frac{-2l}{p_{sgn(-y)}^2(1-e_+-e_-)^2}}$, training with ℓ_{PL} on $x \in X_{S=l}$ improves the individual excessive generalization error* $\mathsf{err}_l^+(\mathcal{P}, \mathcal{A}, x|S)$ *by:*

$$\Omega\left(\frac{l^2}{n^2}\textit{weight}\left(\pi, \left[\frac{2}{3}\frac{l-1}{n-1}, \frac{4}{3}\frac{l}{n}\right]\right) \cdot \sum_{k\neq y}\tilde{\mathbb{P}}[\tilde{y}=k|x]\right)$$

Peer loss extremizes h's prediction to the wrong label with substantial probability for instances with small l: Similar to Theorem 7, when l is small, the power of peer loss does seem to drop:

Theorem 9. *W.p. at least $\frac{1}{\sqrt{2l}}e^{-l \cdot D_{\mathrm{KL}}\left(\frac{1}{2}||e_{sgn(y)}\right)}$, predicting $\mathbb{P}[h(x)=-y]=1$ leads to smaller training loss in ℓ_{PL}.*

$\mathbb{P}[h(x)=-y]=1$ implies that peer loss extremizes h's prediction to the opposite of the true label, and therefore reduces h's generalization power.

REFERENCES

Arpit, D., Jastrzębski, S., Ballas, N., Krueger, D., Bengio, E., Kanwal, M. S., Maharaj, T., Fischer, A., Courville, A., Bengio, Y., et al. A closer look at memorization in deep networks. *arXiv preprint arXiv:1706.05394*, 2017.

Chen, P., Liao, B. B., Chen, G., and Zhang, S. Understanding and utilizing deep neural networks trained with noisy labels. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1062–1070. PMLR, 09–15 Jun 2019. URL http://proceedings.mlr.press/v97/chen19g.html.

Cheng, H., Zhu, Z., Li, X., Gong, Y., Sun, X., and Liu, Y. Learning with instance-dependent label noise: A sample sieve approach. *arXiv preprint arXiv:2010.02347*, 2020a.

Cheng, J., Liu, T., Ramamohanarao, K., and Tao, D. Learning with bounded instance-and label-dependent label noise. In *Proceedings of the 37th International Conference on Machine Learning*, ICML '20, 2020b.

Feldman, V. Does learning require memorization? a short tale about a long tail. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pp. 954–959, 2020.

Feldman, V. and Zhang, C. What neural networks memorize and why: Discovering the long tail via influence estimation. *arXiv preprint arXiv:2008.03703*, 2020.

Han, B., Niu, G., Yu, X., Yao, Q., Xu, M., Tsang, I., and Sugiyama, M. Sigua: Forgetting may make learning with noisy labels more robust. In *International Conference on Machine Learning*, pp. 4006–4016. PMLR, 2020.

Liu, T. and Tao, D. Classification with noisy labels by importance reweighting. *IEEE Transactions on pattern analysis and machine intelligence*, 38(3):447–461, 2016.

Liu, Y. The importance of understanding instance-level noisy labels. *arXiv preprint arXiv:2102.05336*, 2021.

Liu, Y. and Guo, H. Peer loss functions: Learning from noisy labels without knowing noise rates. In *Proceedings of the 37th International Conference on Machine Learning*, ICML '20, 2020.

Lukasik, M., Bhojanapalli, S., Menon, A. K., and Kumar, S. Does label smoothing mitigate label noise? *arXiv preprint arXiv:2003.02819*, 2020.

Menon, A., Van Rooyen, B., Ong, C. S., and Williamson, B. Learning from corrupted binary labels via class-probability estimation. In *International Conference on Machine Learning*, pp. 125–134, 2015.

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

Neyshabur, B., Bhojanapalli, S., McAllester, D., and Srebro, N. Exploring generalization in deep learning. In *Advances in neural information processing systems*, pp. 5947–5956, 2017.

Patrini, G., Rozza, A., Krishna Menon, A., Nock, R., and Qu, L. Making deep neural networks robust to label noise: A loss correction approach. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

Song, H., Kim, M., Park, D., and Lee, J.-G. Prestopping: How does early stopping help generalization against label noise? 2020.

Yao, Q., Yang, H., Han, B., Niu, G., and Kwok, J. T.-Y. Searching to exploit memorization effect in learning with noisy labels. In *International Conference on Machine Learning*, pp. 10789–10798. PMLR, 2020.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

Zhu, X., Anguelov, D., and Ramanan, D. Capturing long-tail distributions of object subcategories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 915–922, 2014.