OT CLEANER: LABEL CORRECTION AS OPTIMAL TRANSPORT

Jun Xia[1,2⋆], *Cheng Tan*[2⋆], *Lirong Wu*[2], *Yongjie Xu*[2], *Stan Z. Li*[2]

[1] Zhejiang University, Hangzhou, 310058, China
[2]AI Lab, School of Engineering, Westlake University, Hangzhou, 310024, China

ABSTRACT

Datasets with noisy labels present challenges for training Deep Neural Networks (DNNs) with high generalization ability. An direct idea is to correct the noisy labels for robust learning. However, existing label correction methods can not handle with heavy noise or datasets with samples of many categories so well. We explain the reasons and introduce a global label distribution regularization to remedy these deficiencies. With this regularization, we convert the label correction to the Optimal Transport (OT) formulation and propose to utilize a fast version of the Sinkhorn-Knopp algorithm for finding an approximate solution efficiently at scale. Experiments on benchmark datasets with both synthetic and real-world label noise show that the superiority of our OT Cleaner in terms of both training efficiency and classification accuracy. The code is available at: `https://github.com/junxia97/OT-Cleaner`.

***Index Terms*—** Image recognition, label noise, label correction, optimal transport

1. INTRODUCTION

Deep learning has demonstrated its superiority in various applications. However, it is extremely expensive and time-consuming to collect large datasets with human annotated labels. To address this problem, there are alternative and inexpensive methods for mining large-scale data with labels, such as querying commercial search engines [1] and downloading social media images with tags [2]. However, these methods are prone to produce incorrect labels. As is shown in a recent research [3], an intractable problem is that Deep Neural Networks (DNNs) can easily over-fit to noisy labels, which dramatically degrades the generalization performance of DNNs. Therefore, it is necessary and urgent to design some valid methods to address this issue.

Previous work showed that during training, DNNs tend to learn simple patterns first, then gradually memorize all samples [4], which justified the widely used small-loss trick: the loss value of clean samples are more likely to be small. Based on this, several existing works select small-loss samples as clean ones to train the DNNs robustly [5]. They achieved significant performance improvements over regular training. Although these methods exclude unreliable samples with the small-loss trick, they may eliminate numerous useful ones among the large-loss samples. Therefore, for a more robust training on noisy labels, SELFIE [6] proposes to refurbish a portion of large-loss samples. However, SELFIE assigns the DNN's prediction as label directly similar to [7],

$$y^{new} = y^{pred}, \tag{1}$$

where y^{pred} is the average of DNN's prediction of several epochs during training and y^{new} is the label after correction respectively. When we train the DNNs with y^{new} as the labels, a trivial global optimal solution will be obtained where a network that always predicts constant label for each sample under high levels of noise. This dilemma has been observed in previous works [8, 7]. To overcome this issue, they add the regularization that labels after correction should be evenly distributed in each batch. However, the samples in each batch are usually class unbalanced. What is worse, this approximation can not handle with datasets with samples of many categories (e.g., CIFAR-100 with 100 classes) so well because the classes number will be near or even larger than the batch size. The samples in each batch can not cover all classes and thus the distribution regularization for each batch can not work well. The other way is to clean the labels with a convex combination of the given noisy label and the prediction of DNNs,

$$y^{new} = \alpha\widehat{y} + (1-\alpha)y^{pred}, \tag{2}$$

where $\alpha \in [0, 1]$ is the label confidence of the given label $\widehat{y}$. Although they avoid the trivial solution, they suffer from the difficulty of accurate estimation of α because α varies across different samples and training stages [9, 10, 8, 11], which results in severe false corrections especially when handling with heavy noise or datasets with samples of many categories.

To tackle these issues, OT Cleaner first selects some small-loss samples for training and then corrects the noisy labels of large-loss samples. In case that the network that always predicts constant label for each sample, OT Cleaner introduces the global label distribution regularization that the labels after correction should be evenly distributed among all classes for all training data instead of each mini-batch (can also be distributed as the noisy label distribution for imbalanced datasets).

'⋆' denotes equal contribution, Stan Z. Li is the corresponding author.

With this regularization, we find the label correction can be formulated as optimal transport [12] and utilize a fast version of Sinkhorn-Knopp algorithm [13] to solve it efficiently. Also, OT Cleaner does not require estimation of label confidence and thus alleviates the severe false correction. To conclude, our key contributions are:

- We explain the poor performance of existing label correction methods when handling with heavy noise and datasets with samples of many categories.
- We introduce a global label distribution regularization to remedy above deficiencies of existing label correction methods. Besides, we formulate label correction as optimal transport and propose a fast version of the Sinkhorn-Knopp algorithm for finding an approximate solution efficiently at scale.
- Experiments on datasets with both synthetic and real-world label noise show that OT Cleaner can achieve competitive performance in terms of training efficiency and classification accuracy.

2. METHOD

2.1. Problem Statement

We consider the problem of K-class classification task using a DNN with a softmax output layer. Suppose that $\mathcal{X} \subset \mathbb{R}^d$ is the input feature space and $\mathcal{Y} = \{0,1\}^K$ be the ground-truth label space in one-hot manner. In a typical classification problem, we are provided a training set $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ obtained from a joint distribution over $\mathcal{X} \times \mathcal{Y}$. The goal is to learn the mapping function $\mathcal{F}(\cdot;\theta)$, a DNN parameterized by θ, to convert the input into a vector of class scores. Then the class scores are mapped to class probabilities with a softmax operator:

$$s(\theta, x_i) = softmax(\mathcal{F}(x_i;\theta)), \tag{3}$$

We optimize θ by minimizing the average cross entropy loss:

$$\mathcal{L}(\theta) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K} y_{i,j}\log s_j(\theta, x_i). \tag{4}$$

In this paper, we focus on learning with noisy labels. Namely, we are provided with a noisy training dataset $\widehat{\mathcal{D}} = \{(x_i, \widehat{y_i})\}_{i=1}^N$ obtained from a noisy joint distribution over $\mathcal{X} \times \mathcal{Y}$, where $\widehat{y_i}$ is a noisy label which may not be true. The label matrix $\widehat{Y} = [\widehat{y_1}, ..., \widehat{y_N}] \in \mathbb{R}^{N\times K}$. Our goal is to mitigate the adverse effects of noisy labels to ensure DNNs generalize well on test data.

2.2. Our method

Let $\mathcal{D}_m$ be the mini-batch samples set. For each batch, we first select some clean samples $\mathcal{D}_s$ based on the widely used small-loss trick [5, 6, 14]. Following previous work [6], we treat $(1-\tau)\times 100\%$ of small-loss samples as clean ones, where τ is the noise rate. If τ is unknown, we can also infer it with cross-validation as previous methods [15, 6]. We denote the left large-loss samples as $\mathcal{D}_l$ and we will substitute their labels with refurbished labels during training. Generally speaking, the part of label correction in our method can be described as follows. We first initialize the label matrix $\widetilde{Y}$ to be optimized with the given unclean label matrix $\widehat{Y}$, i.e., $\widetilde{Y} = \widehat{Y}$. Then we alternate model learning and label correction as bi-level optimization. More specifically, we first learn the model (θ) to obtain the prediction matrix $P \in \mathbb{R}^{N\times K}$ whose (i,j)-th entry is $-\log s_j(\theta, x_i)$ with $\widetilde{Y}$ as the regular training of DNNs. Then we refurbish the label matrix $\widetilde{Y}$ with optimal transport. We consider label correction as an optimization problem where we attempt to minimize the training loss. With the new labels matrix $\widetilde{Y} \in \mathbb{R}^{N\times K}$ whose (i,j)-th entry is $\widetilde{y}_{i,j}$, we can then rewrite Eq. (4) as:

$$\mathcal{L}(\theta, \widetilde{Y}) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K} \widetilde{y}_{i,j}\log s_j(\theta, x_i). \tag{5}$$

In the case of minimizing Eq. (5) with Eq. (1), we will obtain a trivial global optimal solution that the network always predicts constant label for any training sample. To overcome this problem, we add the regularization that the assignments of all labels should be evenly distributed among K classes as previous works about label-noise learning [7, 8]. However, they approximate the label distribution of each class in training data with the label distribution in each mini-batch, which is sub-optimal as analysed in the introduction. Instead, we add this regularization on the total training set:

$$\min_{\theta, \widetilde{Y}} \mathcal{L}(\theta, \widetilde{Y}), \tag{6}$$

$$s.t.\ \forall j : \widetilde{y}_{i,j} = \{0,1\}, \sum_{i=1}^{N}\widetilde{y}_{i,j} = \frac{N}{K};\ \forall i : \sum_{j=1}^{K}\widetilde{y}_{i,j} = 1. \tag{7}$$

We can then formulate label correction as optimal transport problem after relaxing $\widetilde{Y}$:

$$\min_{\widetilde{Y}\in\mathbb{R}_+^{N\times K}} \left\langle \widetilde{Y}, P \right\rangle_F, \tag{8}$$

$$s.t. \quad \widetilde{Y}\mathbb{1}_K = \mathbb{1}_N, \quad \widetilde{Y}^\top \mathbb{1}_N = \frac{N}{K}\cdot \mathbb{1}_K, \tag{9}$$

where $\left\langle \widetilde{Y}, P \right\rangle_F = -\sum_{i,j}\widetilde{y}_{i,j}\log s_j(\theta, x_i)$ is Frobenius inner product of matrices. Note that $\left\langle \widetilde{Y}, P \right\rangle_F = N \times \mathcal{L}(\theta, \widetilde{Y})$. We omit N here because it is a constant for particular dataset.

This is a linear programming which can be solved through simplex algorithm or interior point techniques with computational complexity $\mathcal{O}(N^3 log\, N)$ [16]. In label correction, however, we have to handle with large-scale datasets (e.g.

Algorithm 1 OT Cleaner

Input: $\widehat{\mathcal{D}} = \{(x_i, \widehat{y}_i)\}_{i=1}^{N}$,max epochs E, K, ϵ, $\mathcal{F}(\cdot;\theta)$, warm up epochs E_w, max iterations of sinkhorn algorithm I, optimization schedule $T_o = \{t_k\}_{k=1}^{T}$, T is the optimization times during training.
Output: model parameters θ_T.
Initialize θ_0, $\widetilde{Y}_0 = \widehat{Y}$.
for $t = 0, 1, 2, ..., E-1$ **do**
  for $j = 1$ **to** $\frac{|\widehat{\mathcal{D}}|}{|\mathcal{D}_m|}$ **do**
    if $t < E_w$ **then** ▷ *Warm up*
      $\theta_{t+1} = \theta_t - \alpha\nabla(\frac{1}{|\mathcal{D}_m|}\sum_{x\in\mathcal{D}_m}\mathcal{L}(x, \hat{y}; \theta_t))$
    else
      $\mathcal{D}_s \leftarrow (1-\tau)\times 100\%$ of small-loss samples in $\mathcal{D}_m$; ▷ *Sample Selection*
      $\mathcal{D}_l = \mathcal{D}_m \setminus \mathcal{D}_s$;
      $\theta_{t+1} = \theta_t - \alpha\nabla(\frac{1}{|\mathcal{D}_m|}(\sum_{x\in\mathcal{D}_s}\mathcal{L}(x, \hat{y}; \theta_t)) + \sum_{x\in\mathcal{D}_l}\mathcal{L}(x, \widetilde{y}_t; \theta_t)))$; ▷ *Model training*
    end if
  end for
  if $t \notin T_o$ **then** $\widetilde{Y}_{t+1} = \widetilde{Y}_t$;
  else ▷ *Label correction*
    $P \leftarrow \{-log(softmax(\mathcal{F}(x_i; \theta_{t+1})))\}_{i=1}^{N}$;
    $M = e^{-\frac{P}{\epsilon}}$, $v = \frac{\mathbb{1}_K}{K}$, $l = 0$;
    while $l \leq I$ and not converge **do**
      $u = \frac{\mathbb{1}_N}{Mv}$,$v = \frac{N}{K}\cdot\frac{\mathbb{1}_K}{M^\top u}$; ▷ *Sinkhorn's iteration*
    end while
    $\widetilde{Y}_{t+1} = diag(u)Mdiag(v)$
  end if
end for

Clothing1M) with millions of samples. The computational complexity of simplex algorithm or interior point techniques limits their application in such scale. To address this issue, we resort to the entropy-regularized optimal transport problem:

$$\min_{\widetilde{Y}\in\mathbb{R}_+^{N\times K}} \left\langle \widetilde{Y}, P \right\rangle_F + \epsilon H(\widetilde{Y}), \tag{10}$$

$$s.t. \quad \widetilde{Y}\mathbb{1}_K = \mathbb{1}_N, \quad \widetilde{Y}^\top\mathbb{1}_N = \frac{N}{K}\cdot\mathbb{1}_K, \tag{11}$$

where $H(\widetilde{Y}) = \sum_{i,j}\widetilde{y}_{i,j}\log\widetilde{y}_{i,j}$ is the entropy regularization and $\epsilon > 0$ is the regularization parameter. Adding an entropy regularization to optimal transport problem is computationally more friendly, since it allows the usage of first-order algorithms and it can well approximate the original optimal transport problem with a small enough ϵ [13]. Let $\mathcal{L}_f$ be the Lagrangian function of Eq. (10) and Eq. (11):

$$\begin{aligned}\mathcal{L}_f = \left\langle \widetilde{Y}, P \right\rangle_F + \epsilon H(\widetilde{Y}) - \xi^\top(\widetilde{Y}\mathbb{1}_K - \mathbb{1}_N) \\ -\zeta^\top(\widetilde{Y}^\top\mathbb{1}_N - \frac{N}{K}\cdot\mathbb{1}_K),\end{aligned} \tag{12}$$

where ξ and ζ are dual variables. The KKT condition implies that the optimal solution can be formulated using the optimal dual variables ξ^* and ζ^* as [17],

$$\widetilde{Y}^* = diag(u^*)Mdiag(v^*), \tag{13}$$

$$u^* = e^{\frac{\xi^*}{\epsilon}}, M = e^{-\frac{P}{\epsilon}}, v^* = e^{\frac{\zeta^*}{\epsilon}}, \tag{14}$$

here exponent is element-wise. We can obtain u^* and v^* with Sinkhorn's fixed point iteration,

$$u^{(l+1)} = \frac{\mathbb{1}_N}{Mv^{(l)}}, v^{(l+1)} = \frac{N}{K}\cdot\frac{\mathbb{1}_K}{M^\top u^{(l+1)}}, \tag{15}$$

where the division is entrywise and $v^{(0)} = \frac{\mathbb{1}_K}{K}$. We can then obtain the new label matrix $\widetilde{Y}$ for training samples by:

$$\widetilde{y}_{i,j} = \begin{cases} 1, & \text{if} \quad argmax_{1\leq j\leq K}\widetilde{Y}^*_{i,j} = j \\ 0, & \text{otherwise.}\end{cases} \tag{16}$$

The time complexity of these steps is $\mathcal{O}(NK)$. It is neither $\mathcal{O}(N^2K)$ nor $\mathcal{O}(NK^2)$ because the diagonal matrix exists in Eq. (13). The time complexity scales linearly with the number of samples and thus OT Cleaner can be applied in large-scale datasets. In practice, our algorithm also demonstrates its superiority over competing methods when it comes to training time. Algorithm 1 shows all the steps of our OT Cleaner.

3. EXPERIMENTS

3.1. Experimental settings

Datasets and settings. Our experiments are conducted on three benchmark datasets: CIFAR-10, CIFAR-100 and Clothing1M [18]. For CIFAR datasets, we consider *symmetric* noise which is generated by flipping labels in each class randomly to incorrect labels of other classes. We also evaluate our method on various noise rate τ, with $\tau \in \{20\%, 50\%, 80\%\}$. Besides, Clothing1M consists of 1 million training images collected from online shopping websites with labels generated from surrounding texts. Thus, it has been widely adopted in evaluating algorithms in real-world setting [8, 19].

Baselines. We compare our OT Cleaner with the following label correction methods: (1) Basemodel, which refers to train on noisy datasets with cross entropy loss directly; (2) Bootstrap [9]; (3) F-correction [20]; (4) M-correction [8]; (5) D2L [10]; (6) SELFIE [6]; (7) AdaCorr [19], which is the most recent label correction method.

Implementation details. We utilize ResNet-18 for CIFAR-10 and CIFAR-100 datasets. Besides, Adam optimizer (momentum=0.9) is utilized with an initial learning rate of 0.001, and the batch size is set to 128. We run 200 epochs in total and linearly decay learning rate to zero from 80 to 200 epochs. The aforementioned settings are the same as previous works [5, 21] for fair. For initial convergence of the algorithm, we "warm up" the model for 20 epochs by training on all training data using the standard cross-entropy loss. For Clothing1M, we use ResNet-50 with ImageNet pretrained weights.

Table 1. Average test accuracy (±std in 4 runs) over the last ten epochs on CIFAR-10 and CIFAR-100 with symmetric noise ranging from 20% to 80%. The best and the second best results are highlighted in **bold** and ***italic bold*** respectively.

Datasets		CIFAR-10			CIFAR-100		
Methods/Noise rate		0.2	0.5	0.8	0.2	0.5	0.8
Basemodel		84.81 ± 0.24	61.49 ± 0.58	28.98 ± 0.26	57.79 ± 0.44	33.75 ± 0.46	8.46 ± 0.22
Label correction methods	Bootstrap	86.90 ± 0.40	82.49 ± 0.32	50.28 ± 0.25	58.49 ± 0.13	52.05 ± 0.23	19.89 ± 1.61
	F-correction	87.44 ± 0.15	83.1 ± 0.80	52.16 ± 0.78	60.25 ± 0.10	52.24 ± 0.27	20.64 ± 0.58
	M-correction	89.73 ± 0.12	***84.25 ± 0.19***	***53.93 ± 1.21***	67.26 ± 0.15	***57.25 ± 0.18***	22.69 ± 2.36
	D2L	85.13 ± 0.21	82.37 ± 0.35	50.46 ± 0.65	62.20 ± 0.41	56.98 ± 0.15	***26.75 ± 0.35***
	SELFIE	89.07 ± 0.35	83.67 ± 0.24	51.32 ± 0.48	66.82 ± 0.17	55.67 ± 1.21	25.32 ± 0.85
	AdaCorr	***91.00 ± 0.31***	83.06 ± 0.47	49.33 ± 0.82	**67.77 ± 0.21**	57.12 ± 0.63	24.6 ± 1.13
OT Cleaner		**91.40 ± 0.14**	**85.43 ± 0.29**	**56.93 ± 0.34**	***67.38 ± 0.24***	**58.86 ± 0.13**	**31.20 ± 0.85**

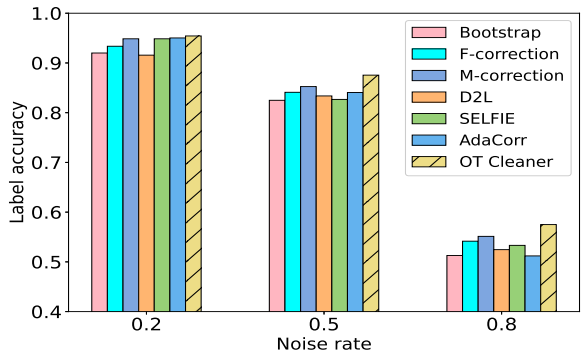

Fig. 1. Label accuracy after correction of various label correction methods on CIFAR-10 with symmetric noise.

3.2. Results on symmetric label noise

The classification accuracies under symmetric label noise on CIFAR-10 and CIFAR-100 are reported in Table 1. It is obvious that our OT Cleaner outperforms other baselines. The superior performance is more pronounced when the noise rates are extremely high. As shown in Figure 1,we also compare OT Cleaner with other label correction methods by evaluating their label correction ability. We can draw the conclusion that OT Cleaner is better at correcting the noisy labels than existing methods, especially under high noise rate and datasets with more categories.

3.3. Results on large-scale dataset with real-world noise

We compare OT Cleaner with existing label correction based methods on Clothing1M with real-world label noise. Note that some previous works conduct experiments by sampling a class-balanced training subset in each epoch (e.g., current state-of-the-art method Dividemix [22]), while others train the model on the full training set [20].To compare with state-of-the-art method and keep fair simultaneously, we thus conduct

Table 2. Test accuracy (mean±std in 3 runs) on Clothing1M.

Training sampling	Standard		Noisy-class-balanced	
Method	Test accuracy	*Training time*	Test accuracy	*Training time*
Basemodel	68.94	-	71.12 ± 0.32	*2.61 ± 0.08h*
F-correction	69.84	-	71.28 ± 0.27	*2.74 ± 0.05h*
M-correction	71.05 ± 0.15	*11.17 ± 0.43h*	-	-
AdaCorr	71.23 ± 0.26	*9.68 ± 0.28h*	-	-
Co-teaching	70.19 ± 0.28	*15.06 ± 0.33h*	72.14 ± 0.28	*4.51 ± 0.07h*
Dividemix	-	-	**73.81 ± 0.41**	*18.78 ± 0.32h*
OT Cleaner	**71.82 ± 0.22**	*12.36 ± 0.31h*	73.38 ± 0.15	*5.69 ± 0.09h*

experiments on both settings and report the results respectively. For standard sampling setting, our OT Cleaner adopts the regularization that the labels after correction must be distributed as the noisy label distribution. For noisy-class-balanced setting, we randomly sample 18976 instances per class for all baselines and adopt the equipartition regularization. The results can be seen in Table 2, which shows that OT Cleaner outperforms other label correction methods. We also present the average training time. Although OT Cleaner falls behind of current state-of-the-art method Dividemix [22] in noisy-class-balanced setting, it is more efficient.

4. CONCLUSION

In this paper, we propose OT Cleaner for robust training against label noise. More specifically, we first select small-loss samples for training and then refurbish large-loss samples with optimal transport. The experimental results demonstrate significant performance gain over competing label correction methods. For the future work, we plan to explore more effective methods for instance-dependent label noise.

5. ACKNOWLEDGMENTS

This work is supported by the Science and Technology Innovation 2030 - Major Project (No. 2021ZD0150100) and National Natural Science Foundation of China (No. U21A20427).

6. REFERENCES

[1] Wen Li, Limin Wang, Wei Li, Eirikur Agustsson, and Luc Van Gool, "Webvision database: Visual learning and understanding from web data," *arXiv preprint arXiv:1708.02862*, 2017.

[2] Dhruv Mahajan, B. Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and van der Laurens Maaten, "Exploring the limits of weakly supervised pretraining," *ECCV*, pp. 185–201, 2018.

[3] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals, "Understanding deep learning requires rethinking generalization," *arXiv preprint arXiv:1611.03530*, 2016.

[4] Devansh Arpit, Stanisław Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al., "A closer look at memorization in deep networks," *arXiv preprint arXiv:1706.05394*, 2017.

[5] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama, "Co-teaching: Robust training of deep neural networks with extremely noisy labels," in *Advances in neural information processing systems*, 2018, pp. 8527–8537.

[6] Hwanjun Song, Minseok Kim, and Jae-Gil Lee, "Selfie: Refurbishing unclean samples for robust deep learning," in *International Conference on Machine Learning*, 2019, pp. 5907–5915.

[7] Daiki Tanaka, Daiki Ikami, Toshihiko Yamasaki, and Kiyoharu Aizawa, "Joint optimization framework for learning with noisy labels," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5552–5560.

[8] Eric Arazo, Diego Ortego, Paul Albert, Noel E O'Connor, and Kevin McGuinness, "Unsupervised label noise modeling and loss correction," *arXiv preprint arXiv:1904.11238*, 2019.

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

[10] Xingjun Ma, Yisen Wang, Michael E Houle, Shuo Zhou, Sarah M Erfani, Shu-Tao Xia, Sudanthi Wijewickrema, and James Bailey, "Dimensionality-driven learning with noisy labels," *arXiv preprint arXiv:1806.02612*, 2018.

[11] Jun Xia, Haitao Lin, Yongjie Xu, Lirong Wu, Zhangyang Gao, Siyuan Li, and Stan Z. Li, "Towards robust graph neural networks against label noise," 2021.

[12] Cédric Villani, *Optimal transport: old and new*, vol. 338, Springer Science & Business Media, 2008.

[13] Marco Cuturi, "Sinkhorn distances: Lightspeed computation of optimal transport," *Advances in neural information processing systems*, vol. 26, pp. 2292–2300, 2013.

[14] Hongxin Wei, Lei Feng, Xiangyu Chen, and Bo An, "Combating noisy labels by agreement: A joint training method with co-regularization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13726–13735.

[15] Tongliang Liu and Dacheng Tao, "Classification with noisy labels by importance reweighting," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 38, no. 3, pp. 447–461, 2015.

[16] Ofir Pele and Michael Werman, "Fast and robust earth mover's distances," in *2009 IEEE 12th International Conference on Computer Vision*. IEEE, 2009, pp. 460–467.

[17] Richard Sinkhorn and Paul Knopp, "Concerning nonnegative matrices and doubly stochastic matrices," *Pacific Journal of Mathematics*, vol. 21, no. 2, pp. 343–348, 1967.

[18] Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang, "Learning from massive noisy labeled data for image classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2691–2699.

[19] Songzhu Zheng, Pengxiang Wu, Aman Goswami, Mayank Goswami, Dimitris Metaxas, and Chao Chen, "Error-bounded correction of noisy labels," in *International Conference on Machine Learning*. PMLR, 2020, pp. 11447–11457.

[20] Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen 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*, 2017, pp. 1944–1952.

[21] Cheng Tan, Jun Xia, Lirong Wu, and Stan Z. Li, *Co-Learning: Learning from Noisy Labels with Self-Supervision*, p. 1405–1413, Association for Computing Machinery, New York, NY, USA, 2021.

[22] Junnan Li, Richard Socher, and Steven CH Hoi, "Dividemix: Learning with noisy labels as semi-supervised learning," *arXiv preprint arXiv:2002.07394*, 2020.