Wasserstein Distributional Normalization For Robust Distributional Certification of Noisy Labeled Data

Sung Woo Park [1] **Junseok Kwon** [1]

Abstract

We propose a novel Wasserstein distributional normalization method that can classify noisy labeled data accurately. Recently, noisy labels have been successfully handled based on small-loss criteria, but have not been clearly understood from the theoretical point of view. In this paper, we address this problem by adopting distributionally robust optimization (DRO). In particular, we present a theoretical investigation of the distributional relationship between uncertain and certain samples based on the small-loss criteria. Our method takes advantage of this relationship to exploit useful information from uncertain samples. To this end, we normalize uncertain samples into the robustly certified region by introducing the non-parametric Ornstein-Ulenbeck type of Wasserstein gradient flows called Wasserstein distributional normalization, which is cheap and fast to implement. We verify that network confidence and distributional certification are fundamentally correlated and show the concentration inequality when the network escapes from over-parameterization. Experimental results demonstrate that our non-parametric classification method outperforms other parametric baselines on the Clothing1M and CIFAR-10/100 datasets when the data have diverse noisy labels.

1. Introduction

The success of deep neural networks in supervised classification tasks is heavily dependent on accurate and high-quality label information. Nevertheless, annotating large-scale datasets is an extremely expensive and time-consuming task. Thus, most conventional studies obtain large-scaled training data using crowd-sourcing platforms (Yu et al., 2018), which inevitably results in *noisy labels* in the annotated samples.

[1]School of Computer Science and Engineering, Artificial Intelligence Graduate School, Chung-Ang University, Seoul, Korea. Correspondence to: Sung Woo Park <pswkiki@gmail.com>, Junseok Kwon <jskwon@cau.ac.kr>.

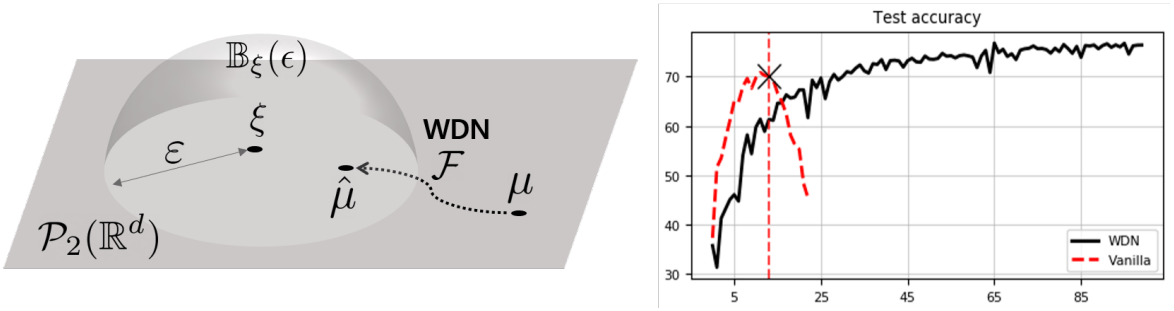
Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

While there are several methods that can deal with noisy-labeled data, recent methods typically adopt the *small-loss* criterion, which helps to construct classification models that are not susceptible to noise corruption. In this learning scheme, a classification network is trained using easy samples first during the early stages of training. Gradually, as training proceeds, more complex samples are selected to train mature models. For example, collaborative learning models (Jiang et al., 2018) have been proposed, in which a mentor network delivers the data-driven curriculum loss to a student network. Dual networks (Han et al., 2018; Yu et al., 2019) generate gradient information jointly using easy samples and then employ this information to allow the networks to teach each other. A disagreement strategy (Wei et al., 2020) updates the gradient information based on disagreement values between dual networks. Accumulated gradients (Han et al., 2020) have been used to escape optimization processes from over-parameterization and obtain more generalized results. These methods have given empirical insight regarding network behavior under noisy labels. However, they have not extensively investigated theoretical aspects of noisy labels.

In contrast, we provide theoretical explanations to understand the network behavior under noisy labels. In particular, we present an in-depth analysis of small-loss criteria in a probabilistic sense. For this, we exploit stochastic properties of noisy-labeled data and develop probabilistic descriptions of data under the small-loss criteria, as follows. Let $\mathbb{P}$ be a probability measure for pre-softmax logits of training samples, l be an objective function for classification, and $\mathbf{1}_{\{\cdot\}}$ be an indicator function. Then, we define *uncertain* and *certain samples* defined in the pre-softmax feature space (*i.e.*, X and Y, respectively), as follows.

$$X \sim \mu|\zeta = \frac{\mathbf{1}_{\{X;l(X)>\zeta\}}\mathbb{P}}{\mathbb{P}[l(X)>\zeta]}, \quad Y \sim \xi|\zeta = \frac{\mathbf{1}_{\{X;l(Y)\leq\zeta\}}\mathbb{P}}{\mathbb{P}[l(Y)\leq\zeta]}, \tag{1}$$

where μ and ξ denote the probability measures of uncertain and certain samples, respectively, and ζ is a constant. In (1), X and Y are sampled from $\mu|\zeta$ and $\xi|\zeta$, respectively. Intuitively, uncertain samples X can be considered as samples that produce losses larger than pre-fixed scalar ζ in each mini-batch. While previous works focus on the usage of Y and the sampling strategy of ζ, they do not thoroughly inves-



(a) Robust distributional certification of μ (b) Classification accuracy

Figure 1. **Distributional certification of μ with the proposed Wasserstein distributional normalization (WDN).** The proposed WDN distributionally certifies uncertain samples by normalizing μ onto the robust certified region $\mathbb{B}_\xi(\varepsilon)$ (the black dotted line in (a)), which forces certified measure $\hat{\mu} = \mathcal{F}[\mu]$ to be in the robust region. In this case, test accuracy is consistently enhanced as training proceeds (the black line in (b)), while an uncertified approach suffers from over-parameterization problems (the red dotted line in (b)).

tigate poor generalization capabilities due to the abundance of uncertain samples X, which potentially contain important information. To exploit the information from the uncertain samples X, we adopt distributionally robust optimization (DRO) and distributionally normalize the uncertain measure μ into $\hat{\mu}$ via transformation $\mathcal{F}$ (*i.e.*, $\hat{\mu} = \mathcal{F}[\mu]$). Then, our objective function is defined as

$$\inf_{\theta} \sup_{\hat{\mu}} \mathbb{E}_{\hat{X}\sim\hat{\mu}}[l(\hat{X},\theta)] + \mathbb{E}_{Y\sim\xi}[l(Y,\theta)], \tag{2}$$

where $\hat{X}$ denotes transformed uncertain samples (*i.e.*, certified samples) and θ indicates learnable parameters for classification networks. Specifically, the transformation $\mathcal{F}$ is designed to satisfy the following inequality:

$$\sup_{\hat{\mu}} \mathcal{W}(\hat{\mu},\xi) \leq \varepsilon, \tag{3}$$

where $\mathcal{W}(\hat{\mu},\xi)$ indicates the statistical discrepancy (*i.e.*, Wasserstein distance) between the normalized measure $\hat{\mu}$ and certain measure ξ. In this setting, our network can be trained using both the certified samples $\hat{X} \sim \hat{\mu} = \mathcal{F}[\mu]$ and certain samples $Y \sim \xi$. Fig.1 shows the distributional certification of μ with our method.

If a network is trained based on uncertified approaches (*e.g.*, conventional cross-entropy) and noisy signals are consecutively provided, the network quickly suffers from over-parameterization problems owing to the inconsistency of noisy-labeled training data. To prevent this problem, we determine a certified robust region $\mathbb{B}_\xi$ centered at ξ, which contains a certified measure $\hat{\mu} \in \mathbb{B}_\xi$ normalized from μ. To empirically verify the effectiveness of DRO settings in (2) for noisy-labeled data, we derive a well-posed upper bound ε (Section 5), which helps the proposed network to escape from over-parameterization. In addition, we develop universal concentration inequality (Section 6), which captures the probabilistic state when our network is robust to noisy-labeled data. The explicit form and theoretical/numerical advantages of $\mathcal{F}$ are presented (Section 4). Please note that our distributional normalization method is **fully non-parametric, simple, and computationally efficient**. Thus, our method can reduce the computational complexity of conventional approaches for dual networks, while maintaining the concept of small-loss criterion.

Main contributions of our work are as follows.

- We theoretically verify that there exists a *strong correlation* between model confidence and statistical distance between X and Y. We empirically investigate that the classification accuracy worsens when uncertified samples are consecutively given to classification networks.
- We develop a *simple, non-parametric*, and *computationally efficient* stochastic model to control the observed ill-behaved sample dynamics. For this, we present Wasserstein gradient flows of uncertain measure and simulate non-parametric stochastic differential equations (*i.e.*, Ornstein-Ulenbeck process) for tractable computation. Thus, our method requires no additional learning parameter.
- We provide two important theoretical results. 1) the exponentially controllable certification bound ε is introduced, which makes our method to control the distributional certification. 2) the concentration inequality of certified measure is presented, which clearly describes the probabilistic resemblance between $\hat{\mu}$ and ξ.

2. Related Work

Curriculum Learning & Small-loss Criterion. To handle noisy labels, (Han et al., 2018; Yu et al., 2019; Jiang et al., 2018; Wei et al., 2020; Lyu & Tsang, 2020a; Han

et al., 2020) adopted curriculum learning or sample selection frameworks. However, these methods considered only a small number of selected samples, while a large number of samples are excluded at the end of the training. While this inevitably leads to poor generalization capabilities, conventional sample selection methods cannot solve this problem, because a large number of training samples are gradually eliminated. (Chen et al., 2019) iterated cross-validation for randomly partitioned noisy-labeled data to identify samples with correct labels. To generate the partitions and select samples, they adopted the small-loss criteria. In contrast, our method can extract useful information from unselected samples $X \sim \mu$ (*i.e.*, uncertain samples) by distributionally certified samples ($e.g.$, $\hat{X} \sim \mathcal{F}[\mu]$) for accurate classification.

Loss Correction & Label Correction. Noisy labels were transformed either explicitly or implicitly into clean labels by correcting classification losses (Patrini et al., 2017b; Hendrycks et al., 2018; Ren et al., 2018). While modifying loss-dynamics failed to accurately correct noisy labels under extremely noisy environments, (Arazo et al., 2019) adopted a label augmentation method called MixUp (Zhang et al., 2018). Unlike these methods, our method transforms holistic information from uncertain samples into certain samples, which implicitly reduces effect of potentially noisy labels.

Distillation. To mitigate the impact of gradients induced by noisy labels, (Li et al., 2019b) updated mean teacher parameters by calculating an exponential moving average of student parameters. (Lukasik et al., 2020) deeply investigated the effect of label smearing for noisy labels and linked label smoothing to loss correction in a distillation framework. Like these methods, our method leverages useful properties of distillation models.

Other methods. A robust generative classifier based on pre-trained deep models has been proposed (Lee et al., 2019). (Damodaran et al., 2019) designed a constraint on the Wasserstein space and adopted an adversarial framework for classification models of noisy-labeled data by implementing semantic Wasserstein distance. (Pleiss et al., 2020) identified noisy-labeled samples by considering AUM statistics, which exploit the differences in the training dynamics of clean and mislabeled samples. Recently, (Li et al., 2019a) adopted semi-supervised learning methods to deal with noisy labels, where the student network utilized labeled and unlabeled samples to perform semi-supervised learning guided by the teacher network.

3. Distributional Robust Optimization

Let l be the conventional cross-entropy loss and $\hat{r}$ be a corrupted label for an unknown label transition matrix from a clean label r with label transition matrix Q. Then, a conventional objective function for classification with noisy labels can be defined as follows:

$$\inf_{\theta} \mathcal{J}[\mu] = \inf_{\theta} \mathbb{E}_{X\sim\mu,\hat{r}|Q}\left[l(X;\theta,\hat{r})\right]. \tag{4}$$

However, as aforementioned in Section 1, the conventional objective function defined in (4) cannot be used for accurate classification if the network is under noisy-labeled data. Instead of abandoning uncertain samples $X \sim \mu$ as in previous works, we normalize μ in the certified region $\mathbb{B}_\xi$ and the network uses the information of certified samples $\hat{\mu}$ for accurate classification. For a clear mathematical description, we first introduce the following definition.

Definition 1. *(Wasserstein certified region) Let $\mathcal{P}_2(\mathbb{R}^d)$ be a 2-Wasserstein space. We define a Wasserstein certified region in this space as follows:*

$$\mathbb{B}_\xi(\varepsilon) = \left\{\mu \in \mathcal{P}_2\left(\mathbb{R}^d\right) : \mathcal{W}(\mu,\xi) \le \varepsilon\right\}, \tag{5}$$

where $\mathcal{W}$ denotes the 2-Wasserstein distance.

Then, we propose a distributionally robust (DR) objective function as follows:

$$\begin{aligned}&\inf_{\theta} \sup_{\hat{\mu}\in\mathbb{B}_\xi(\varepsilon)} \mathcal{J}[\hat{\mu}] + \mathcal{J}[\xi] \\ &= \inf_{\theta} \mathbb{E}_{\hat{X}\sim\mathcal{F}[\mu],\hat{r}}[l(\hat{X};\theta,\hat{r})] + \mathbb{E}_{X\sim\xi,\hat{r}}[l(Y;\theta,\hat{r})],\end{aligned} \tag{6}$$

where $\mathcal{F} : \mathcal{P}_2(\mathbb{R}^d) \to \mathcal{P}_2(\mathbb{R}^d)$ is a distributional normalization for probability measure μ, which ensures that $\mathcal{F}[\mu]$ is lying in the certified region (*i.e.*, $\hat{\mu} = \mathcal{F}[\mu] \in \mathbb{B}_\xi(\varepsilon)$).

As the normalization functional $\mathcal{F}$ is assumed to satisfy the condition in (3), the sup operation can be omitted in the objective function in (6). Then, the probability measure ξ is defined as $\xi = \arg\min \mathcal{J}\left[\xi_{k^\star}\right]$, where ξ_k denotes a certain measure at the k-th iteration and $k^\star \in \mathbb{I}_{k-1} = \{1, \cdots, k-1\}$. In other words, ξ indicates the best certain measure that produces the smallest losses so far at training time. Thus, our objective function is formulated to train the network using certain samples Y as well as certified uncertain samples $\hat{X}$ under noisy labels $\hat{r}$.

Adaptive Radius of Certified Region. An important question arises from the objective function in (6): *How can we select the effective radius ε of the certified region $\mathbb{B}_\xi$?* To answer this question, (Sinha et al., 2018) postulate the empirical Lagrangian relaxation regarding Monge map of optimal transport to ensure the radius of certified region: $\mathcal{W}(\xi,\hat{\mu}) \le \varepsilon$. However, in this setting (*i.e.*, robustness to adversarial examples), ε cannot be designed, but needs to be given in advance according to various noisy environments. Please note that in our setting (*i.e.*, robustness to noisy labels), there is no unified rule or prior information on selecting an effective radius ensuring accurate classification.

As a naive solution, one may search the optimal certification radius ε by manually tuning this value according to various network architectures, datasets, and noise types. However, because a large amount of computation is required, this approach is computationally inefficient. To overcome this, we propose an adaptive radius ε that is data-dependent and cheap to compute. In particular, we decompose the original certification radius ε into two separate terms $(\varepsilon_1, \varepsilon_2)$ by using the triangle inequality of the Wasserstein distance:

$$\begin{aligned}\mathcal{W}(\xi, \hat{\mu}) &\leq \varepsilon = \varepsilon_1 + \varepsilon_2 \\ &= \underbrace{\mathcal{W}(\xi, \mathcal{N}(\mathbf{m}_\xi, \Sigma_\xi))}_{\varepsilon_1\text{: Intrinsic statistics}} + \underbrace{\mathcal{W}(\mathcal{N}(\mathbf{m}_\xi, \Sigma_\xi), \hat{\mu})}_{\varepsilon_2\text{: Distributional Normalization}}. \end{aligned} \quad (7)$$

Intrinsic Statistics. The first term (*i.e.*, ε_1) sets a detour point as a Gaussian measure, for which the mean and covariance are the same as those for ξ (*i.e.*, $\mathbf{m}_\xi = \mathbb{E}_{Y\sim\xi}[Y]$ and $\Sigma_\xi = \mathbf{Cov}_{Y\sim\xi}[Y]$). The Wasserstein distance of this term is only dependent on the statistical structure of ξ because $(\mathbf{m}_\xi, \Sigma_\xi)$ is dependent on ξ. Thus, this term can induce a *data-dependent and non-zero constant* certification radius, whenever $\xi \neq \mathcal{N}$ and can prevent ε from collapsing to $\varepsilon \to 0$. This provides a marked advantages compared to manual tuning of certification radius ε, because this term is automatically determined based on the characteristics of noise types, datasets, and network architectures during training. Furthermore, owing to the independence between ε_1 and ε_2, a divide-and-conquer-based analyses can be conducted to theoretically investigate the properties of ε.

Distributional Normalization. The second term (*i.e.*, ε_2) represents our central objective to design. $\mathcal{F}$ facilitates the distributional normalization of μ to certify the geometric conditions. Based on the independence of ε_1 and ε_2, the original condition in (3) can be rewritten as follow.

$$\begin{aligned}&\sup_{\hat{\mu}} \mathcal{W}(\xi, \hat{\mu} = \mathcal{F}[\mu]) \\ &\leq \varepsilon_1 + \sup_{\hat{\mu}} \mathcal{W}(\mathcal{N}(\mathbf{m}_\xi, \Sigma_\xi), \hat{\mu} = \mathcal{F}[\mu]) = \varepsilon_1 + \varepsilon_2. \end{aligned} \quad (8)$$

In the following section, we introduce the explicit form of $\mathcal{F}$, which ensures the certification radius in (8), and demonstrate its theoretical/numerical advantages.

4. Wasserstein Distributional Normalization

In the previous section, we propose a DR-type objective function that performs distributional normalization $\mathcal{F}$ such that the normalized measure $\mathcal{F}[\mu]$ lies in $\mathbb{B}_\xi(\varepsilon)$ for the certified radius ε. In this section, we specify the formulation of $\mathcal{F}$ and show the theoretical and numerical advantages.

Definition 2. *The functional $\mathcal{F} : \mathbb{R}^+ \times \mathcal{P}_2 \to \mathcal{P}_2$ on the probability measure such that $\mathcal{F}_t[\mu] = \mu_t$ is called as **distributional normalization** if μ_t is a solution to the following continuity equations:*

$$\partial_t \mu_t = \nabla \cdot (\mu_t v_t)^1, \quad (9)$$

where $d\mu_t = p_t d\mathcal{N}_\xi, d\mathcal{N}_\xi = dq_t dx$. For simplicity, we denote $\mathcal{N}_\xi = \mathcal{N}(\mathbf{m}_\xi, \Sigma_\xi)$.

The distributional normalization $\mathcal{F}$ is defined as a gradient flow in the Wasserstein space beginning at μ. While the steady state of the solution μ_t (*e.g.*, $t \to \infty$) is a Gaussian measure, the following property is satisfied:

Proposition 1. *The distributional normalization $\mathcal{F}$ maps μ into the certified robust region with controllable radius $\varepsilon_2 = K_2(\mu)e^{-t}$ (i.e., $\mathbb{B}_{\mathcal{N}_\xi}(K_2 e^{-t}(\mu))$), where $K_2(\mu) > 0$ is a constant that depends on μ.*

It is well known that the solution to (9) induces a geodesic in the Wasserstein space (Villani, 2008), which is the shortest path from $\mu = \mu_{t=0}$ to $\mathcal{N}_\xi$. The functional $\mathcal{F}_t$ generates a path for μ_t, in which the distance is exponentially decayed according to the auxiliary variable t and constant K_2, meaning $\mathcal{W}(\mathcal{N}_\xi, \mathcal{F}_t\mu) \leq \varepsilon_2 = K_2 e^{-t}$. This theoretical results indicates that the Wasserstein distance in (8) can be *controlled*. Thus, by setting a different t, our method can efficiently control the distance. Specifically, the certification radius in (8) can be written to a controllable form:

$$\sup_{\hat{\mu}} \mathcal{W}(\xi, \hat{\mu}) \leq \varepsilon_1 + \sup_{\hat{\mu}} \mathcal{W}(\mathcal{N}_\xi, \hat{\mu}) = \varepsilon_1 + K_2 e^{-t}. \quad (10)$$

However, it is typically intractable to directly compute the continuity in (9). To solve this problem, we adopt stochastic differential equations (SDEs), which enable tractable computation. In particular, we adopt an *Ornstein-Ulenbeck (OU) process*[2], which can be approximated using particle-based dynamics. We draw i.i.d $N(1-\varrho)$ uncertain samples (*i.e.*, $\{X^n_{t=0}\} \sim \mu$) from a single batch with N samples using (1) for hyper-parameter $0 \leq \varrho \leq 1$. Then, we simulate a discrete SDE for each particle using the Euler-Maruyama scheme:

$$X^n_{t+1} = X^n_t - \nabla\phi(X^n_t; \mathbf{m}_\nu)\Delta_t + \sqrt{2\tau^{-1}\Delta_t \Sigma_\xi} Z^n, \quad (11)$$

where $\phi(X_t; \mathbf{m}_\nu) = \frac{\tau}{2} d_E^2(X_t, \mathbf{m}_\nu)$, $n \in \{1 \cdots, N(1-\varrho)\}$, Z is a standard Gaussian random variable, d_E is Euclidean distance, and N is a mini-batch size. Thus, the simulated samples are indicated as follows:

$$\{X^n_t\}_{n \leq N(1-\varrho)} = \{\hat{X}^n\}_{n \leq N(1-\varrho)} \sim \hat{\mu} = \mathcal{F}_t[\mu] \quad (12)$$

By the property of OU process, the proposed method is computationally efficient and non-parametric to estimate

[1] This equation is read as $\partial_t p_t(x) = \nabla \cdot (p_t(x) \nabla \log q_t(x))$ in a distributional sense, where $v_t = \nabla \log q_t$.

[2] The corresponding partial differential equation (PDE) is the Fokker-planck equation $\partial_t p_t(x) = \nabla \cdot (p_t(x) \nabla \log q_t(x))$ defined in (9).

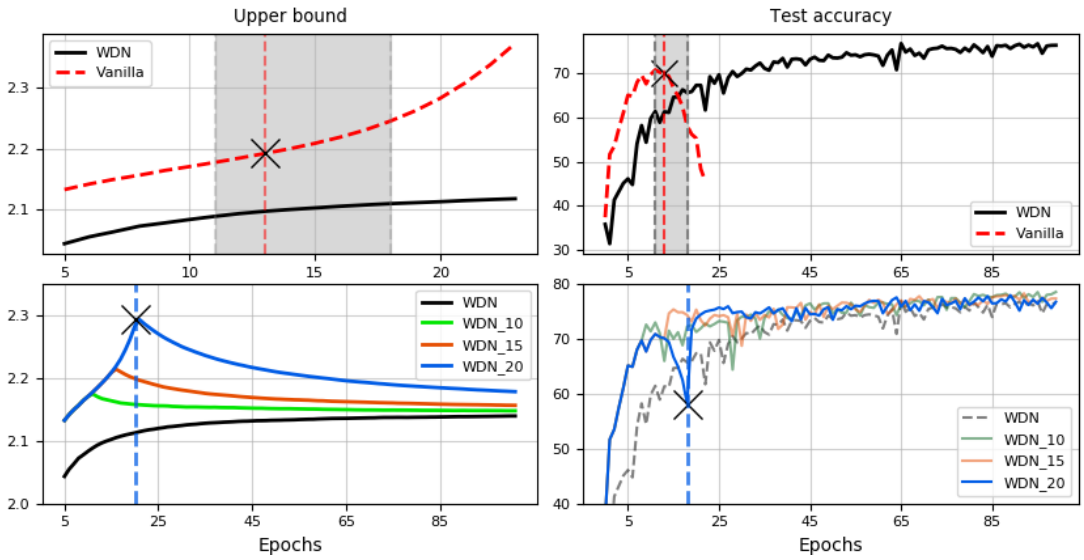


Figure 2. **Relation between the certification radius ε and test accuracy**.

the second term ε_2 in (7), because the SDE in (11) for the OU process has a simple form with fixed drift (*i.e.*, $\nabla\phi$) and diffusion (*i.e.*, $\sqrt{\Sigma_\xi}Z$) terms. Because these terms are independent to times, we can induce an non-parametric simulation of the SDE in (11) and make our method *computationally efficient* compared to other model-based methods that require additional dual networks (Han et al., 2018).

Explicit certification radius. In addition to the aforementioned motivation on setting a detour point as a Gaussian measure, decomposing ε into $\varepsilon_1 + \varepsilon_2$ in (7) provides theoretical advantages (*i.e.*, explicit certification radius). The next proposition investigates the explicit certification radius regarding two terms $(\varepsilon_1, \varepsilon_2)$ in (7), which are induced by the property of (9). Let define $v_t(x) = \nabla \log p_t(x)$ where $\frac{d\mu_t}{d\mathcal{N}_\xi} = p_t(x)$ is a distribution as defined in (2). Let $\mathcal{L}$ be a differential operator such that $\mathcal{L}[f] = \Sigma_\xi \nabla^2[f] - \mathbf{m}_\xi^T \nabla f$ [3] which acts on the space of compactly supported continuous function (*i.e.*, $f \in C_0^\infty$). Then, we define the integral operators K_2 on $\mathcal{P}_2$ as follows:

$$K_2(\mu) = \sqrt{\sup_f \int_{\mathbb{R}^d} |\mathcal{L}f(x)|\, d\mu(x)}. \tag{13}$$

Intuitively, $K_2(\mu)$ can be interpreted as an indicator that tells us how the uncertain measure μ is diffused according to Gaussian measure. For example, if $\mu = \mathcal{N}_\xi$ in (13), then $K_2(\mathcal{N}_\xi) = 0$. By using property of integral operator K_2, an explicit certification radius can be found.

Proposition 2. *Let $0 < \beta(\xi) < \infty$ be a numerical constant, which depends on ξ. Then, the following inequality holds:*

$$\mathcal{W}(\xi, \hat{\mu}) \le \varepsilon = \varepsilon_1 + \varepsilon_2 = K_1(\xi) \vee \left[e^{-t} K_2(\mu) + K_2(\xi)\right], \tag{14}$$

where $\lambda_{\max}(\Sigma_\xi)$ denotes the maximum eigenvalue of the covariance matrix Σ_ξ and $K_1(\xi) = \sqrt{d\beta\lambda_{max}(\Sigma_\xi)} + \|\mathbb{E}_\xi Y\|_2$ which is only dependent on ξ.

Finally, we obtain the controllable *explicit certification radius* in (14). By setting a fixed large t, the certification radius ε is adaptively determined based on different characteristics of datasets and network architectures.

5. Empirical Observations

In this section, we investigate the effectiveness of DRO with the proposed WDN regarding the certified region $\mathbb{B}_\xi(\varepsilon)$. For this, we numerically measure the certification radius ε developed in (14) and demonstrate that distributional certification makes classification networks robust to noisy labels.

As the first term $K_1(\xi)$ in (14) is constant and typically very small compared to the second term with a large t, we only examine the behavior of the second term $K_2(\xi) + K_2(\mathcal{F}_t\mu)$, which can be efficiently estimated. In particular, for any probability measure ν, $K_2(\nu)$ can be estimated[4]:

$$K_2(\nu) \approx \mathbb{E}_{X,Z} \left[\left\| e^{-\Delta} X + \sqrt{1 - e^{-2\Delta}} (\Sigma_\nu^{\frac{1}{2}} Z + \mathbf{m}_\nu) \right\| - \|X\|\right], \tag{15}$$

where Z denotes the standard Gaussian random variable. In this paper, we set $\Delta = 0.01$ and $t \in [16, 64]$. As we aforementioned, $K_2(\mathcal{N}_\xi) = 0$.

[3] ∇ and ∇^2 indicate gradient and hessian operators, respectively.

[4] Please refer to Supplementary Materials for detailed mathematical descriptions.

Algorithm 1 Wasserstein Distributional Normalization (**WDN**)

Require: $\alpha \in [0, 0.2]$, $\varrho \in [0.1, 0.65], T \in [16, 64], \Delta_t = 10^{-4}, \tau \in [0.1, 0.001]$,
  for $k = 1$ to K (*i.e.*, the total number of training iterations) **do**
    1) Select $(1-\rho)N$ uncertain and ρN certain samples from the mini-batch N.
    $\{Y_k^n\}_{\{n \le \rho N\}} \sim \xi_k, \{X_k^n\}_{\{n \le (1-\rho)N\}} \sim \mu_k$
    2) Update the certain measure ξ.
    if $\mathcal{J}[\xi_k] < \mathcal{J}[\xi]$ **then**
      $\xi \leftarrow \xi_k, \mathbf{m}_\xi \leftarrow \mathbb{E}[Y_k]$, and $\Sigma_\xi \leftarrow \mathbf{Cov}[Y_k]$
    end if
    3) Update the moving geodesic average $\mathcal{N}(\mathbf{m}^\alpha, \Sigma^\alpha)$.
    Solve the Ricatti equation $\mathcal{T}\Sigma_\xi\mathcal{T} = \Sigma_{\xi_k}$.
    $\Sigma^\alpha = ((1-\alpha)\mathbf{I}_d + \alpha\mathcal{T})\,\Sigma_\xi\,((1-\alpha)\mathbf{I}_d + \alpha\mathcal{T})$ and $\mathbf{m}^\alpha = (1-\alpha)\mathbf{m}_\xi + \alpha\mathbf{m}_{\xi_k}$
    4) Simulate the discrete SDE for T steps.
    for $t = 0$ to $T-1$ **do**
      $X_{k,t+1}^n = -\nabla\phi(X_{k,t}^n; \mathbf{m}^\alpha)\Delta_t + \sqrt{2\tau^{-1}\Sigma_\xi^\alpha}dW_t^n$ *s.t.* $\{X_{k,t=0}^n\} \sim \mu_k, \{X_{k,t=T}^n\} \sim \mathcal{F}_T\mu_k$
    end for
    5) Update the network using the objective function.
    $\mathcal{J}[\mathcal{F}[\mu_k]] + \mathcal{J}[\xi_k] = \mathbb{E}_{\mathcal{F}_T\mu_k}[l(X_{k,T}; \theta, \hat{r})] + \mathbb{E}_{\xi_k}[l(Y_k; \theta, \hat{r})]$
  end for

Using (15), we observe three important empirical properties:

(1) Uncertified samples induce low test accuracy. We examine the relationship between certification radius ε and test accuracy in an experiment using the CIFAR-10 dataset with symmetric noise at a ratio of 0.5. Fig.2 presents the landscape for the $\log_{10}$-scaled cumulative average of certification radius ε and test accuracy over epochs. The red dotted lines represent the landscape of $\varepsilon = \mathcal{W}(\xi, \mu)$ for the uncertified approach with cross-entropy loss. The black lines indicate the landscape of certification radius for the certified network with proposed WDN, where $\sup_{\hat{\mu}} \mathcal{W}(\xi, \hat{\mu}) = \varepsilon_k = \varepsilon_1 + \varepsilon_2 = K_2(\xi_k) + K_2(\mathcal{F}_{t=T}\mu_k)$. Note that the test accuracy of the classification network with uncertified samples begins to decrease after 13-epochs (red-dotted vertical lines in the top-right plot), whereas the certification radius increases quadratically in the top-left plot. These experimental results verify that *uncertified samples inevitably induce low test accuracy*.

Contrary to the uncertified network, if the distributional certification is ensured by WDN, the certification radius can be efficiently controlled (*i.e.*, $\limsup_k \varepsilon_k \approx 2.15$). In this case, the test accuracy continues to increase, even after 13-epochs. For detailed analysis, we compute the deviations as follows: $\hat{\Delta}_k = \varepsilon_k - \varepsilon_{k-1}$. In the gray regions, the deviation for the uncertified network is grater than 2.5×10^{-2}, *i.e.*, $\Delta_k > 2.5 \times 10^{-2}$. Then, its test accuracy begins to drop, as shown in Fig.2. In contrast to the uncertified network, the maximum deviation of the certification radius is very small ($\sup_k \hat{\Delta}_k \le 8 \times 10^{-3}$) if distributional certification is ensured by the proposed WDN.

(2) The distributional certification with our WDN helps networks to escape from over-parameterization. To analyze the behavior of classification network under over-parameterization with and without the distributional certification, we design several variants of experiments, which begin at delayed epochs. The green, orange, and blue curves in the second row of Fig.2 represent the landscapes, when our WDN is applied after $k_d \in \{10, 15, 20\}$ epochs, respectively. In this case, the certification radius ε_k is

$$\varepsilon_k = \begin{cases} \mathcal{W}(\xi_k, \mu_k), & \text{if } k < k_d, \text{ uncertified,} \\ K_2(\xi_k) + K_2(\mathcal{F}_{t=T}\mu_k), & \text{else } k \ge k_d, \text{ certified.} \end{cases} \tag{16}$$

Consider $k_d = 20$, which is described by the blue dotted vertical lines. Before our WDN is applied (*i.e.*, $k < k_d$), the network suffers from over-parameterization, which induces a significant performance drop, as shown in the blue curve of the bottom-right plot. However, the network rapidly recovers to normal accuracy if distributional certification is assured by the normalization (*i.e.*, $k \ge k_d$). Note that similar behavior can be observed in the green and orange curves. In particular, the orange curve produces fewer fluctuations than the blue curve in terms of test accuracy. This shows that our WDN can help networks to escape from over-parameterization by certifying distributional robustness.

(3) The certification radius ε with our WDN is dependent of data statistics. Another interesting point in Fig.2 is that all curves, excluding the red curve, converge to particular numbers $2.15 = \underline{\varepsilon} := \liminf_k \varepsilon_k \le \limsup_k \varepsilon_k := \bar{\varepsilon} = 2.2$. The upper bound $\bar{\varepsilon}$ is neither overly enlarged nor collapsed to zero, while the lower bound $\underline{\varepsilon}$ is fixed for all curves. We argue that this behavior stems from the distributional characteristics of the proposed method, where the first term in (7), $\mathcal{W}(\xi, \mathcal{N}_\xi) \propto K_2(\xi)$, is a non-zero data-dependent term that is minimized by the proposed method. Therefore, we can derive the following relationship:

$$\begin{aligned} [\mathcal{W}(\xi, \hat{\mu}) \le \mathcal{W}(\xi, \mathcal{N}_\xi) + \mathcal{W}(\mathcal{N}_\xi, \mathcal{F}\mu)] \Downarrow \\ \propto\ [K_2(\xi) + K_2(\mathcal{F}\mu) = \varepsilon] \Downarrow. \end{aligned} \tag{17}$$

This empirical observation verifies that a detour point, which is set as a Gaussian measure, *can induce the data-dependent bound* $(\underline{\varepsilon}, \bar{\varepsilon})$, where our data-dependent bound can change according to various noise levels and efficiently leverage data-dependent statistics. Fig.2 indicates that classification models with more stable certification also induce more stable convergence in test accuracy.

6. Probabilistic Concentration

In Section 5, we investigate the effectiveness of WDN to certify distributional robustness on noisy-labeled data where uncertain measure μ is normalized into the certified region $\mathbb{B}_\xi(\varepsilon)$. In this circumstance, the following proposition indicates the probabilistic resemblance of ξ and certified $\hat{\mu}$.

Proposition 3. *There exists $\delta > 0$ such that the following concentration inequality for an uncertain measure holds:*

$$\hat{\mu}\big(\,|\sigma - \mathbb{E}_\nu[\sigma]| \geq \delta\big) \leq 6e^{-\frac{\sqrt{2}\delta^{\frac{3}{2}}}{K_2(\mu)}}, \tag{18}$$

where σ denotes the soft-max function.

In (18), we show that the network inference using the certified measure is similar to that of certain measure $\mathbb{E}_\nu[\sigma]$, where the upper-bound in right-hand side is exponentially relative to the initial diffuseness of μ (*i.e.*, $K_2(\mu)$), which induces long-tail probabilistic representations. This indicates that the proposed WDN certifies uncertain measures to make $\hat{\mu}$ similar to ξ.

7. Wasserstein Moving Geodesic Average

In the experiments, we observed that the certain measure $\xi = \arg\min \mathcal{J}\left[\xi_{k^\star}\right]$ was not updated for a few epochs after the training begins. This is problematic because ξ can diverge significantly from the current ξ_k, which is equivalent to the normalized measure $\hat{\mu}_k = \mathcal{F}[\mu_k]$ at epoch k diverging from ξ, meaning $\hat{X} \sim \hat{\mu}_k$ and $Y \sim \xi$ become statistically inconsistent. To alleviate this statistical distortion, we modify the detour measure from $\mathcal{N}_\xi$ to another Gaussian measure, which allows us to capture the statistics of ξ_k and ξ. Inspired by the moving average of Gaussian parameters in batch normalization (Ioffe & Szegedy, 2015), we propose the *Wasserstein moving geodesic average*. Specifically, we replace Gaussian parameters $\{\mathbf{m}_\xi, \Sigma_\xi\}$ with $\{\mathbf{m}^\alpha, \Sigma^\alpha\}$ such that $\mathbf{m}^\alpha = (1-\alpha)\mathbf{m}_\xi + \alpha\mathbf{m}_{\xi_k}$ and $\Sigma^\alpha = ((1-\alpha)\mathbf{I}_d + \alpha\mathcal{T})\,\Sigma_\xi\,((1-\alpha)\mathbf{I}_d + \alpha\mathcal{T})$, where $\mathcal{T}$ is a solution to the Riccati equation $\mathcal{T}\Sigma_\xi\mathcal{T} = \Sigma_{\xi_k}$. Therefore our final detour Gaussian measure is set to $\mathcal{N}_\xi^\alpha := \mathcal{N}(\mathbf{m}(\alpha), \Sigma(\alpha)), 0 \leq \alpha \leq 1$[5]. The overall procedure for our method is summarized in Algorithm 1.

8. Experiments

8.1. Experiments on the CIFAR-10/100 dataset

We used settings similar to those proposed by (Laine & Aila, 2017; Han et al., 2018) for our experiments on the CIFAR10/100 dataset. We used a 9-layered CNN as the baseline architecture with a batch size of 128. We used an Adam optimizer with $(\beta_1, \beta_2) = (0.9, 0.99)$, where the learning rate linearly decreased from 10^{-3} to 10^{-5}.

Synthetic Noise. We injected label noise into clean datasets using a noise transition matrix $Q_{i,j} = \mathbf{Pr}(\hat{r} = j | r = i)$, where a noisy label $\hat{r}$ is obtained from a true clean label r. We defined $Q_{i,j}$ by following the approach discussed by (Han et al., 2018). For symmetric noise, we used the polynomial, $\varrho = -1.11r^2 + 1.78r + 0.04$ for $0.2 \leq r \leq 0.65$, where r is the noise ratio. For asymmetric noise, we set ϱ to 0.35. To select the enhanced detour measure, we set α to 0.2 for the Wasserstein moving geodesic average in all experiments. We trained our classification model over 500 epochs because the test accuracy of our method continued increasing, whereas those of the other methods did not. We compared our method with other state-of-the-art methods, including [MentorNet, (Jiang et al., 2018)], [Co-teaching, (Han et al., 2018)], [Co-teaching+, (Yu et al., 2019)], [GCE, (Zhang & Sabuncu, 2018)], [RoG, (Lee et al., 2019)], [JoCoR, (Wei et al., 2020)], [NPCL, (Lyu & Tsang, 2020b)], [SIGUA, (Han et al., 2020)], and [DivideMix, (Li et al., 2019a)]. As shown in Table 1, the proposed WDN significantly outperformed other baseline methods. Please note that our WDN utilizes OU-process and corresponding Gaussian measure as our main objects, there are potential risks when handling highly concentrated and non-smooth types of noise (*e.g.*, asymmetric noise). Nevertheless, the proposed WDN still produced accurate results, even with asymmetric noise. In this case, a variant of our WDN (*i.e.*, WDN_{cot}) exhibited the best performance.

Open-set Noise. In this experiment, we considered the open-set noisy scenario suggested by (Wang et al., 2018), where a large number of training images were sampled from other CIFAR-100 dataset. However, these images were labeled according to the classes in the CIFAR-10 dataset. We used a 9-layered CNN, which was also used in our previous experiment. For hyper-parameters, we set ϱ and α to 0.5 and 0.2, respectively. As shown in Table 2, our method achieved state-of-the-art accuracy.

Collaboration with Other Methods. Because our core methodology is based on small-loss criteria, our method can collaborate with co-teaching methods. In (Han et al., 2018), only certain samples $(Y \sim \xi)$ were used for updating colleague networks, where the number of uncertain samples gradually decreased until it reached a predetermined value. To enhance potentially bad statistics for co-teaching,

[5] Please refer to Supplementary Materials for more details.

Table 1. **Average test accuracy (%) on the CIFAR-10/100 dataset over the last 10 epochs with various noise corruptions.** The symbol * indicates scores provided by the corresponding authors. WDN$_{cot}$ denotes our WDN combined with a co-teaching network. The best results are boldfaced.

Methods	Symmetric 20%	Symmetric 50%	Asymmetric 45%
Vanilla	$71.91 \pm .43 / 40.44 \pm .36$	$49.54 \pm .41 / 21.34 \pm .27$	$49.06 \pm 1.02 / 31.85 \pm .85$
MentorNet*	$80.76 \pm .36 / 52.13 \pm .40$	$71.10 \pm .48 / 39.00 \pm 1.00$	$58.14 \pm .38 / 31.60 \pm .51$
Co-teaching$^+$	$80.64 \pm .15 / 56.15 \pm .09$	$58.43 \pm .30 / 37.88 \pm .06$	$70.78 \pm .11 / 32.88 \pm .25$
GCE	$84.68 \pm .05 / 51.86 \pm .09$	$61.80 \pm .11 / 37.60 \pm .08$	$61.09 \pm .18 / 33.13 \pm .14$
RoG*	84.32 / 58.16	76.67 / 45.42	71.26 / 43.18
JoCoR	$85.73 \pm .19 / 53.01 \pm .04$	$79.41 \pm .25 / 43.49 \pm .46$	$64.21 \pm .12 / 26.51 \pm .32$
NPCL*	$84.30 \pm .07 / 55.30 \pm .09$	$77.66 \pm .09 / 42.56 \pm .06$	–
SIGUA*	≤ 84 / –	≤ 78 / –	≤ 65 / –
DivideMix	–	$81.13 \pm .18$ / $\mathbf{49.41 \pm .25}$	$68.93 \pm .33$ / $34.24 \pm .63$
WDN	$\mathbf{87.40 \pm .23} / \mathbf{59.18 \pm .29}$	$\mathbf{82.89 \pm .13} / 48.45 \pm .27$	$\mathbf{76.12 \pm .29} / 38.23 \pm .31$
Co-teaching	$78.23 \pm .27 / 53.89 \pm .09$	$72.81 \pm .20 / 34.96 \pm .50$	$70.46 \pm .58 / 34.55 \pm .12$
WDN$_{cot}$	$87.12 \pm .16 / 57.27 \pm .33$	$76.06 \pm .28 / 42.38 \pm .28$	$74.11 \pm .35 / \mathbf{44.41 \pm .37}$

Table 2. **Test accuracy on the CIFAR-10 dataset with open-set noisy labels from CIFAR-100**.

Methods	Vanilla	GCE	Co-teaching	Co-teaching$^+$	JoCoR	WDN
Accuracy	38.12	46.57	35.77	42.57	47.73	**51.28**

Table 3. **Test accuracy (mean, %) on the Clothing 1M dataset.**

Methods	GCE	D2L	FW	JoCoR	WAR	SL	JOFL	DMI	MLNT	PENCIL	WDN	DivideMix
Accuracy	69.0	69.47	69.84	70.30	70.66	71.02	72.23	72.46	73.47	73.49	74.75	74.76

Table 4. **Average training time for the 5-epochs (sec) on the CIFAR-10 dataset.**

Methods	Vanilla	GCE	WDN	Co-teaching	JoCoR	DivideMix
Time	$11.43 \pm .05$	$11.53 \pm .06$	$12.72 \pm .08$	$15.88 \pm .11$	$17.88 \pm .11$	$34.41 \pm .53$
Δ	$+0\%$	$+9\%$	$+11.3\%$	$+38.9\%$	$+56.3\%$	$+201\%$

we taught dual networks by considering a set of samples (Y, X_T), where $X_T \sim \mathcal{F}_T\mu$ are certified samples using (11). Table 1 shows the test accuracy results for the proposed collaboration model with a co-teaching network (WDN$_{cot}$).

8.2. Experiments on a Real-world dataset

To evaluate our method on real-world datasets, we employed the Clothing1M dataset presented by (Xiao et al., 2015), which consists of 1M noisy, labeled, and large-scale cloth images with 14 classes collected from shopping websites. It contains 50K, 10K, and 14K clean images for training, testing, and validation, respectively. We only used a *noisy* set for training; for testing, we used a *clean* set. We set $\alpha = 0.2$ and $\varrho = 0.1$. For fair comparison, we followed the settings suggested in previous works. We used a pre-trained ResNet50 for a baseline architecture with a batch size of 48. For the pre-processing steps, we applied a random center crop, random flipping, and normalization to 224×224 pixels. We adopted the Adam optimizer with a learning rate starting at 10^{-5} that linearly decayed to 5×10^{-6} at 24K iterations. Regarding the baseline methods, we compared the proposed method to [GCE, (Zhang & Sabuncu, 2018)], [D2L, (Ma et al., 2018)], [FW, (Patrini et al., 2017a)], [WAR, (Damodaran et al., 2019)], [SL, (Wang et al., 2019)], [JOFL, (Tanaka et al., 2018)], [DMI, (Xu et al., 2019)], [PENCIL, (Yi & Wu, 2019)], and [MLNT, (Li et al., 2019b)]. Table 3 reveals that our method achieved competitive performance as comparison with other baseline methods.

8.3. Computational Cost

Because Co-teaching, JoCoR, and DivideMix use additional networks, the number of network parameters is twice $(8.86M)$ that of the vanilla network $(4.43M)$. In Table 4, we compared the average training time for the first 5-epochs over various baseline methods under symmetric noise on the CIFAR-10 dataset. While non-parametric methods such as GCE and WDN required less than 12% additional time, other methods that require additional networks spent more time than non-parametric methods. The averaging time can change according to various experimental environments. In Table 4, we measured the time using publicly available code provided by authors.

9. Conclusion

We proposed a novel method called WDN for accurate classification of noisy labels. The proposed method normalizes uncertain measures to robustly certified region by adopting

Wasserstein gradient flow. To this end, we simulated discrete SDE using the Euler-Maruyama scheme, which makes our method fast, computationally efficient, and non-parametric. In theoretical analysis, we derived the explicit certification radius of the proposed Wasserstein normalization and experimentally demonstrated a strong relationship between distributional certification and the over-parameterization. We conducted experiments on the CIFAR-10/100 and Clothing1M datasets. The experimental results demonstrated that the proposed WDN significantly outperforms other state-of-the-art methods.

Acknowledgements

This work was supported by Institute of Information communications Technology Planning Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2021-0-01341, Artificial Intelligence Graduate School Program(Chung-Ang university)).

References

Arazo, E., Ortego, D., Albert, P., O'Connor, N., and Mcguinness, K. Unsupervised label noise modeling and loss correction. In *ICML*, 2019.

Chen, P., Liao, B. B., Chen, G., and Zhang, S. Understanding and utilizing deep neural networks trained with noisy labels. In *ICML*, 2019.

Damodaran, B. B., Fatras, K., Lobry, S., Flamary, R., Tuia, D., and Courty, N. Wasserstein adversarial regularization (WAR) on label noise. *arXiv preprint arXiv:1904.03936*, 2019.

Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., Tsang, I. W., and Sugiyama, M. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *NeurIPS*, 2018.

Han, B., Niu, G., Yu, X., Yao, Q., Xu, M., Tsang, I. W., and Sugiyama, M. Sigua: Forgetting may make learning with noisy labels more robust. In *ICML*, 2020.

Hendrycks, D., Mazeika, M., Wilson, D., and Gimpel, K. Using trusted data to train deep networks on labels corrupted by severe noise. In *NeurIPS*, 2018.

Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015.

Jiang, L., Zhou, Z., Leung, T., Li, L.-J., and Fei-Fei, L. Mentornet: Regularizing very deep neural networks on corrupted labels. In *ICML*, 2018.

Laine, S. and Aila, T. Temporal ensembling for semi-supervised learning. In *ICLR*, 2017.

Lee, K., Yun, S., Lee, K., Lee, H., Li, B., and Shin, J. Robust inference via generative classifiers for handling noisy labels. In *ICML*, 2019.

Li, J., Socher, R., and Hoi, S. C. Dividemix: Learning with noisy labels as semi-supervised learning. In *International Conference on Learning Representations*, 2019a.

Li, J., Wong, Y., Zhao, Q., and Kankanhalli, M. S. Learning to learn from noisy labeled data. In *CVPR*, 2019b.

Lukasik, M., Bhojanapalli, S., Menon, A., and Kumar, S. Does label smoothing mitigate label noise? In *ICML*, 2020.

Lyu, Y. and Tsang, I. W. Curriculum loss: Robust learning and generalization against label corruption. In *ICLR*, 2020a.

Lyu, Y. and Tsang, I. W. Curriculum loss: Robust learning and generalization against label corruption. In *ICLR*, 2020b.

Ma, X., Wang, Y., Houle, M. E., Zhou, S., Erfani, S., Xia, S., Wijewickrema, S., and Bailey, J. Dimensionality-driven learning with noisy labels. In *ICML*, 2018.

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 *CVPR*, 2017a.

Patrini, G., Rozza, A., Menon, A. K., Nock, R., and Qu, L. Making deep neural networks robust to label noise: A loss correction approach. In *CVPR*, 2017b.

Pleiss, G., Zhang, T., Elenberg, E. R., and Weinberger, K. Q. Identifying mislabeled data using the area under the margin ranking. In *NeurIPS*, 2020.

Ren, M., Zeng, W., Yang, B., and Urtasun, R. Learning to reweight examples for robust deep learning. In *ICML*, 2018.

Sinha, A., Namkoong, H., and Duchi, J. Certifiable distributional robustness with principled adversarial training. In *ICLR*, 2018.

Tanaka, D., Ikami, D., Yamasaki, T., and Aizawa, K. Joint optimization framework for learning with noisy labels. In *CVPR*, 2018.

Villani, C. *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 2008. ISBN 9783540710509.

Wang, Y., Liu, W., Ma, X., Bailey, J., Zha, H., Song, L., and Xia, S.-T. Iterative learning with open-set noisy labels. In *CVPR*, 2018.

Wang, Y., Ma, X., Chen, Z., Luo, Y., Yi, J., and Bailey, J. Symmetric cross entropy for robust learning with noisy labels. In *ICCV*, 2019.

Wei, H., Feng, L., Chen, X., and An, B. Combating noisy labels by agreement: A joint training method with co-regularization. In *CVPR*, 2020.

Xiao, T., Xia, T., Yang, Y., Huang, C., and Wang, X. Learning from massive noisy labeled data for image classification. In *CVPR*, 2015.

Xu, Y., Cao, P., Kong, Y., and Wang, Y. L_dmi: A novel information-theoretic loss function for training deep nets robust to label noise. In *NeurIPS*, 2019.

Yi, K. and Wu, J. Probabilistic end-to-end noise correction for learning with noisy labels. In *CVPR*, 2019.

Yu, X., Liu, T., Gong, M., and Tao, D. Learning with biased complementary labels. In *ECCV*, 2018.

Yu, X., Han, B., Yao, J., Niu, G., Tsang, I., and Sugiyama, M. How does disagreement help generalization against label corruption? In *ICML*, 2019.

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. mixup: Beyond empirical risk minimization. In *ICLR*, 2018.

Zhang, Z. and Sabuncu, M. Generalized cross entropy loss for training deep neural networks with noisy labels. In *NeurIPS*, 2018.