Confidence Scores Make Instance-dependent Label-noise Learning Possible

Antonin Berthon [1 2]   Bo Han [3 1]   Gang Niu [1]   Tongliang Liu [4]   Masashi Sugiyama [1 5]

Abstract

In learning with noisy labels, for every instance, its label can randomly walk to other classes following a *transition distribution* which is named a *noise model*. Well-studied noise models are all *instance-independent*, namely, the transition depends only on the original label but not the instance itself, and thus they are less practical in the wild. Fortunately, methods based on instance-dependent noise have been studied, but most of them have to rely on strong assumptions on the noise models. To alleviate this issue, we introduce *confidence-scored instance-dependent noise* (CSIDN), where each instance-label pair is equipped with a confidence score. We find that with the help of confidence scores, the transition distribution of each instance can be approximately estimated. Similarly to the powerful forward correction for instance-independent noise, we propose a novel *instance-level forward correction* for CSIDN. We demonstrate the utility and effectiveness of our method through multiple experiments on datasets with synthetic label noise and real-world unknown noise.

1. Introduction

The recent success of deep neural networks has increased the need for high-quality labeled data. However, such a labelling process can be time-consuming and costly. A compromise is to resort to weakly supervised annotations, using crowdsourcing platforms or trained classifiers that annotate the data automatically. These weakly supervised annotations tend to be low-quality and noisy, which negatively affects the accuracy of high-capacity models due to mem-

[1]RIKEN [2]ENS Paris-Saclay [3]Hong Kong Baptist University [4]University of Sydney [5]University of Tokyo. Correspondence to: Antonin Berthon <berthon.antonin@gmail.com>, Bo Han <bhanml@comp.hkbu.edu.hk>, Gang Niu <gang.niu@riken.jp>, Tongliang Liu <tongliang.liu@sydney.edu.au>, Masashi Sugiyama <sugi@k.u-tokyo.ac.jp>.

orization effects (Zhang et al., 2017). Thus, learning with noisy labels has often drawn a lot of attention.

Early works on noisy labels studied *random classification noise* (RCN) for binary classification (Angluin & Laird, 1988; Kearns, 1993). In the RCN model, each instance has its label flipped with a fixed noise rate $\rho \in [0, \frac{1}{2})$. A natural extension of RCN is *class-conditional noise* (CCN) for multi-class classification (Stempfel & Ralaivola, 2009; Natarajan et al., 2013; Scott et al., 2013; Menon et al., 2015; van Rooyen & Williamson, 2018; Patrini et al., 2016; Yao et al., 2020b; Zhang et al., 2021; Wang et al., 2021) (Appendix A). In the CCN model, each instance from class i has a fixed probability $\rho_{i,j}$ of being assigned to class j. Thus, it is possible to encode some similarity information between classes. For example, we expect that the image of a "dog" is more likely to be erroneously labelled as "cat" than "boat".

To handle the CCN model, a common method is the *loss correction*, which aims to correct the prediction or the loss of the classifier using an estimated noise transition matrix (Patrini et al., 2017; Sukhbaatar et al., 2015; Goldberger & Ben-Reuven, 2017; Ma et al., 2018; Li et al., 2021; Wu et al., 2021). Another common approach is the *label correction*, which aims to improve the label quality during training. For example, (Reed et al., 2015) introduced a bootstrapping scheme. Similarly, (Tanaka et al., 2018) proposed to update the weights of a classifier iteratively using noisy labels, and use the updated classifier to yield more high-quality pseudo-labels for the training set. Although these methods are able to handle the simplified CCN model, they are unable to cope with real-world noise, e.g., *instance-dependent noise* (IDN).

The IDN model considers a more general noise (Manwani & Sastry, 2013; Ghosh et al., 2014; Menon et al., 2018; Cheng et al., 2020; Menon et al., 2018; Scott, 2019; Cheng et al., 2021; Xia et al., 2020; Zhu et al., 2021), where the probability that an instance is mislabeled depends on both its class and features. Intuitively, this noise is quite realistic, as poor-quality or ambiguous instances are more likely to be mislabeled in real-world datasets. However, it is much more complex to formulate the IDN model, since the probability of a mislabeled instance is a function of not only the label space but also the input space that can be very high-dimensional.

As a result, several pioneer works have considered stronger

Table 1. Comparisons between baselines and our work for handling the IDN model. Rate-identifiability denotes whether the transition matrix is identifiable.

Approaches	Multi-class	Rate-identifiability	Unbounded-noise
(Du & Cai, 2015)	✗	✗	✓
(Menon et al., 2018)	✗	✓	✓
(Bootkrajang & Chaijaruwanich, 2018)	✗	✗	✓
(Cheng et al., 2020)	✗	✓	✗
Our work	✓	✓	✓

assumptions on noise functions. However, stronger assumptions tend to restrict the utility of these works (Table 1). For instance, the boundary-consistent noise model considers stronger noise for samples closer to the decision boundary of the Bayes-optimal classifier (Du & Cai, 2015; Menon et al., 2018). However, such a model is restricted to binary classification and cannot estimate noise functions. (Cheng et al., 2020) recently studied a particular case of the IDN model, where noise functions are upper-bounded. Nonetheless, their method is limited to binary classification and has only been tested on small datasets.

Instead of simplifying assumptions on noise functions, we propose to tackle the IDN model from the source, by considering *confidence scores* to be available for the label of each instance. We term this new setting *confidence-scored instance-dependent noise* (CSIDN, Figure 1c). The confidence scores denote how likely an instance is to be correctly labeled. Assuming that (i) confidence scores are available for each instance, (ii) transitions probabilities to other classes are independent of the instance conditionally on the assigned label being erroneous and (iii) a set of anchor points (Liu & Tao, 2015) is available, we derive an *instance-level forward correction* algorithm which can fully estimate the transition probability for each instance, and subsequently train a robust classifier with a loss-correction method similarly to (Patrini et al., 2017).

Our rationale is that in tasks involving instance-dependent noise, the confidence information can be easily derived with no extra cost. Specifically, the confidence information can be available in automatic annotation via a softmax output layer of deep neural networks. This layer outputs an estimate of the probability that each class is observed: when a model outputs a given class with probability 0.9, we expect the predicted class to be true 9 times out of 10 on average.

In theory, when the loss used during training is classification-calibrated (Zhang et al., 2004; Bartlett et al., 2006) and proper composite (Reid & Williamson, 2010; Nock & Nielsen, 2009), the class-posterior probability of the assigned label can be approximately interpreted as a confidence measure that the label is correct. Therefore, for multi-class classification, when training deep neural networks via the cross-entropy loss, the final-layer outputs of deep neural networks can be approximately seen as confidence scores, since the cross-entropy loss is classification-calibrated and proper composite (Gneiting & Raftery, 2007).

To sum up, we first formulate *instance-dependent noise* in Section 2.1, and expose its robustness challenge in Section 2.2. Then, we formally give a definition for confidence scores and propose the *confidence-scored instance-dependent noise* (CSIDN) model in Section 2.3. Meanwhile, to handle this new noise model, we present the first practical algorithm termed *instance-level forward correction* in Section 3, and validate the proposed algorithm through extensive synthetic and real experiments in Section 4. Finally, we conclude in Section 5.

2. Tackling instance-dependent noise from the source

This section presents the IDN model along with the limitations of existing approaches, and introduces the CSIDN model as a tractable instance-dependent noise model.

2.1. Noise models: from class-conditional noise to instance-dependent noise

We formulate the problem of learning with noisy labels here. Let D be the distribution of a pair of random variables $(X, Y) \in \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X} \in \mathbb{R}^d$, $\mathcal{Y} = \{1, 2, \dots, K\}$ and K is the number of classes. In the classification task with noisy labels, we hope to train a classifier while having only access to samples from a noisy distribution $\bar{D}$ of random variables $(X, \bar{Y}) \in \mathcal{X} \times \mathcal{Y}$. Given a point x sampled from X, $\bar{Y}$ is derived from the random variable Y via a noise transition matrix $T(x) = (T_{i,j}(x))_{i,j=1}^K \in [0, 1]^{K \times K}$:

$$P(\bar{Y} = j | X = x) = \sum_{i=1}^K T_{i,j}(x) P(Y = i | X = x). \quad (1)$$

Each noise function $T_{i,j} : \mathcal{X} \mapsto [0, 1]$ is defined as $T_{i,j}(x) = P(\bar{Y} = j | Y = i, X = x)$. In the *class-conditional noise* (CNN) model (Figure 1a), the transition matrix does not depend on the instance x and the noise is entirely characterized by the K^2 constants $T_{i,j}$. However, in the *instance-dependent noise* (IDN) model (Figure 1b), the transition matrix depends on the actual instance.
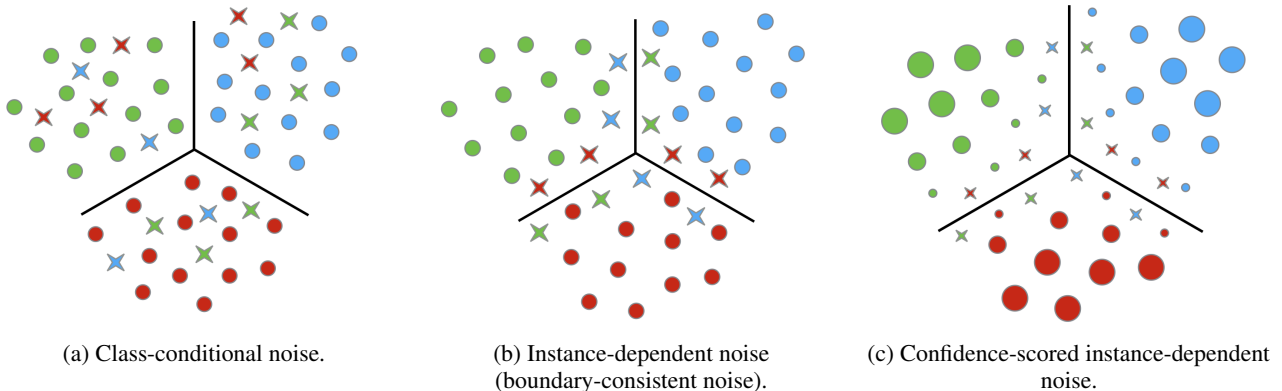
(a) Class-conditional noise.

(b) Instance-dependent noise (boundary-consistent noise).

(c) Confidence-scored instance-dependent noise.

Figure 1. Illustration of different noise models. Each color represents an observed class $\bar{y}$: circles indicate $\bar{y} = y$, while crosses indicate $\bar{y} \neq y$. The size of each point represents the confidence scores in the label $\bar{y}$: the bigger the point is, the more confident the label is. (a) In the CCN model, the noise function only depends on the label of each instance. (b) In the IDN and CSIDN models, the noise function depends on the observed instance x. To illustrate the IDN model, we show a special case called boundary-consistent noise, i.e., points that lie close to the decision boundary are more likely to be mislabelled. (c) The CSIDN model varies from the IDN model in that each instance is associated with a confidence score (Section 2.3).

This tremendously complicates the problem, as the noise is now characterized by K^2 functions over the latent space $\mathcal{X}$, which can be very high-dimensional (e.g., $d \sim 10^4 - 10^6$ for an object recognition dataset (Krizhevsky, 2009; Netzer et al., 2011)).

2.2. Challenges from instance-dependent noise

Limitation of existing CCN methods. Due to the complexity of the IDN model, most recent works in learning with noisy labels have focused on the CCN model (Stempfel & Ralaivola, 2009; Natarajan et al., 2013; Scott et al., 2013; Menon et al., 2015; van Rooyen & Williamson, 2018; Patrini et al., 2016; Han et al., 2018a) (Figure 1a) since the CCN model can be seen as a simplified IDN model (Figure 1b) free of feature information.

In addition to *loss correction* and *label correction* mentioned before, another method for the CCN model is *sample selection*, which aims to find reliable samples during training, such as the small-loss approaches (Jiang et al., 2018; Han et al., 2018b). Inspired by the memorization in deep learning (Arpit et al., 2017; Yao et al., 2020a; Han et al., 2020; Xia et al., 2021), those methods first run a standard classifier on a noisy dataset, then select the small-loss samples for reliable training.

However, all approaches cannot handle the IDN model directly. Specifically, *loss correction* considers the noise model to be characterized by a *fixed* transition matrix, which does not include any instance-level information. Meanwhile, *label correction* is vulnerable to the IDN model, since the classifier will be much weaker on noisy regions and labels corrected by the current prediction would likely be erro-

neous. Similarly, *sample selection* is easily affected by the IDN model.

For example, in the small-loss approaches, instance-dependent noise functions can leave partial regions of the input space clean and other regions very noisy (e.g., in an object recognition dataset, poor-quality pictures will tend to receive more noisy labels than high-quality ones (Netzer et al., 2011)). Since clean regions will tend to receive smaller losses than noisy regions, the small-loss approaches, which only train on points with the smallest losses, will focus on clean regions and neglect harder noisy regions. Then, since the distribution of clean regions will subsequently be different from the global distribution, this will introduce a covariate shift (Sugiyama & Kawanabe, 2012; Shimodaira, 2000), which greatly degrades performances. Moreover, it is hard to use importance reweighting (Sugiyama et al., 2007) to alleviate the issue, since importance reweighting requires estimating the clean posterior probability that is intractable for the IDN model.

To demonstrate this fact, we generate a 3-class distribution of concentric circles (*cf.* Figure 2a), with $\forall (x, y) \in \mathbb{R}^2 \times \{1, 2, 3\}$, $P(\bar{y} \neq y|x) = \frac{1}{2}\left(\frac{w \cdot x}{\|w\|\|x\|} + 1\right)$ with $w = (0, 1)$ (*cf.* Figure 2b). We then train a network on the top $R(T)$ small-loss instances at each epoch T based on the losses of the previous epoch, with $R(T)$ decreasing in T as described in (Han et al., 2018b). Figure 2c shows the density of the top 50% small-loss instances selected after 10 epochs: since noisy regions are associated to higher losses, the network eventually tends to select instances from the clean region and neglect the noisy region. This leads to covariate shift, which results in decreased performances (Shimodaira, 2000).
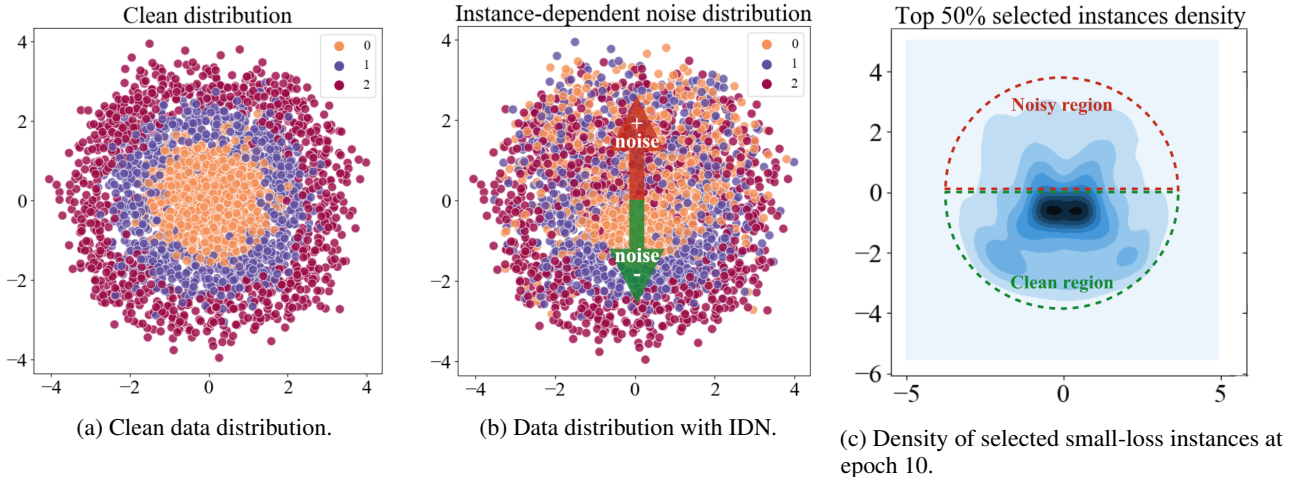
(a) Clean data distribution. (b) Data distribution with IDN. (c) Density of selected small-loss instances at epoch 10.

Figure 2. The limitation of the small-loss approaches in the IDN model. (a) Clean distribution. (b) instance-dependent noise in the direction $w = (0, 1)$ with an average corruption rate of 40%: points towards the upper region are more likely to be corrupted than points towards the bottom region. (c) Density map of the instances selected by a small-loss approach at epoch 10. The sample selection gets biased towards clean regions. Since the clean and noisy regions have different distributions, selecting most instances from clean regions creates a covariate shift between the training and test distributions, which can greatly degrades performances.

Limitation of pioneer IDN methods. The main challenge of the IDN model is the wide range of possible noise functions included in its formulation. Since each $T_{i,j}(\cdot)$ is a function of the high-dimensional input space $\mathcal{X}$, it is challenging for a model to be flexible enough to fit any real-world noise function while being trainable on corrupted datasets, let alone derive theoretical results. Instead, various recent works have considered stronger assumptions on noise functions.

For instance, *boundary-consistent noise* (BCN), first introduced by (Du & Cai, 2015) and generalized in (Menon et al., 2018), considers stronger noise for samples closer to the decision boundary of the Bayes-optimal classifier. This is a reasonable model for noise from human annotators, since "harder" instances (i.e., instances closer to the decision boundary) are more likely to be corrupted. Moreover, it is simple enough to derive some theoretical guarantees, as done in (Menon et al., 2018). Additionally, an extension of the BCN model was studied in (Bootkrajang & Chaijaruwanich, 2018), where the noise function is approximated as a mixture of Gaussians. However, the BCN model and its extension are restricted to binary classification, and their geometry-based assumption becomes difficult to fathom for high-dimensional input spaces.

Furthermore, (Cheng et al., 2020) recently studied a particular case of the IDN model, where the probabilities that the true labels of samples flip into corrupted ones have upper bounds. They proposed a method based on distilled samples, where noisy labels agree with the Bayes-optimal classifier on the clean distribution. However, their method is also limited to binary classification and has only been tested on small UCI datasets. Table 1 summarizes the characteristics of those approaches.

2.3. Confidence-scored instance-dependent noise

Instead of simplifying assumptions on noise functions, we propose to tackle the IDN model from the source. Namely, we consider that, for each instance, we have access to a measure of confidence in the assigned label. As most of noisy datasets arise from crowdsourcing or automatic annotation, such confidence scores can be easily derived during the dataset construction, often with no extra cost. This allows for a good approximation of noise functions with weaker assumptions.

Definition of confidence scores. For any data point $(x, \bar{y})$ sampled from the joint distribution $(X, \bar{Y})$, we define the confidence score r_x as follows.

$$r_x = P(Y = \bar{y} | \bar{Y} = \bar{y}, X = x). \quad (2)$$

Namely, the probability that the assigned label is correct.

CSIDN: A tractable instance-dependent noise model. Recall the intrinsic difficulty of the IDN model: to fully characterize this noise, one would need to estimate K^2 functions $T_{i,j}(\cdot)$ over the input space $\mathcal{X}$. This is of course intractable with a finite noisy dataset. This is why pioneer solutions to the IDN model have been so far limited by very strong assumptions.

However, considering additional confidence scores, one can wonder whether such information would make the IDN model tractable with less restrictive assumptions. Hence, we introduce a new and tractable instance-dependent noise model: *confidence-scored instance-dependent noise* (CSIDN, Figure 1c). In this noise model, the training data takes the form $S := \{(x_i, \bar{y}_i, r_{x_i}), i = 1, \ldots, N\}$, where $\{(x_i, \bar{y}_i)\}_i \overset{\text{i.i.d.}}{\sim} \bar{D}$ and $r_{x_i} = P(Y = \bar{y}_i | \bar{Y} = \bar{y}_i, X = x_i)$ is the previously defined confidence scores in the assigned label of a given instance (Eq. (2)). The confidence information r_x is decisive for robustness to instance-dependent noise, as it provides a proxy for the noise functions $T_{i,j}$ of the training data that are often intractable.

3. Benchmark solution for handling the CSIDN model

To tackle the CSIDN model, we propose a benchmark solution. Inspired by *forward correction* (Patrini et al., 2017) for the CCN model, we want to correct each prediction $P(\bar{y}|x)$ with the noise transition matrix $T(x)$. However, the transition matrix for the CSIDN model is instance-dependent, being estimated for each instance x. We term our solution *instance-level forward correction*.

3.1. Estimating instance-dependent transition matrix

Using the confidence scores, we will first estimate the diagonal terms $(T_{i,i}(\cdot))_{i=1}^K$ of the transition matrix, and then estimate the non-diagonal ones.

Diagonal terms. The diagonal terms of the transition matrix correspond to the probabilities that assigned labels are equal to true labels. However, the confidence scores available are only relevant to the class corresponding to the observed label. Therefore, we need to proceed differently whether the confidence scores are available for the considered class or not.

First, note that for each sample $(x, \bar{y}, r_x) \in S_i := \{(x, \bar{y}, r_x) \in S | \bar{y} = i\}$, $T_{i,i}(x)$ can be derived for the most part from the confidence scores alone: $\forall (x, \bar{y}, r_x) \in S_i$,

$$
\begin{aligned}
T_{i,i}(x) &= P(\bar{Y} = i | Y = i, X = x) \\
&= P(Y = i | \bar{Y} = i, X = x) \frac{P(\bar{Y} = i | X = x)}{P(Y = i | X = x)} \\
&= r_x \, \beta_i(x), \quad\quad (3)
\end{aligned}
$$

where $\beta_i(x) = \frac{P(\bar{Y}=i|X=x)}{P(Y=i|X=x)}$.

In practice, we use an iterative procedure to estimate in turn $\beta_i(\cdot)$ and $T_{i,i}(\cdot)$ (see Section 3.2 for details). Then, for the rest of samples $(x, \bar{y}, r_x) \in S \backslash S_i$, r_x does not give any direct information on $T_{i,i}(\cdot)$. Hence, we simply set each function $T_{i,i}(\cdot)$ as its empirical mean μ_i estimated using

samples from S_i at the current epoch: $\forall (x, y, r_x) \in S \backslash S_i$,

$$
\hat{T}_{i,i}(x) = \frac{1}{|S_i|} \sum_{(x', \bar{y}', r'_x) \in S_i} T_{i,i}(x') = \mu_i, \quad\quad (4)
$$

where $|S|$ denotes the cardinality of S.

Non-diagonal terms. For non-diagonal terms, we have: $\forall i \neq j, \forall x \in \mathcal{X}$,

$$
\begin{aligned}
T_{i,j}(x) &= P(\bar{Y} = j | Y = i, X = x) \\
&= P(\bar{Y} = j, \bar{Y} \neq i | Y = i, X = x) \\
&= P(\bar{Y} = j | \bar{Y} \neq i, Y = i, X = x) \\
&\quad \cdot P(\bar{Y} \neq i | Y = i, X = x) \\
&= \alpha_{i,j}(x)(1 - T_{i,i}(x)), \quad\quad (5)
\end{aligned}
$$

where $\alpha_{i,j}(x) = P(\bar{Y} = j | \bar{Y} \neq i, Y = i, X = x)$.

In Eq. (5), $\alpha_{i,j}(x)$ refers to the probability that an instance x with true label i has an observed label j, once we know that the observed label is different from the true one. Then, a reasonable assumption is that $\forall i \neq j, \forall x \in \mathcal{X}, \alpha_{i,j}(x) = \alpha_{i,j}$: conditionally on the observed label being erroneous, the class transitions are not influenced by the instance x. In other words, the dependence on x of the noise function only impacts the "magnitude" of the noise and not the class transitions.

To illustrate this assumption, consider a crowdsourcing task of object recognition where some classes are very similar to each other and can only be well identified if a particular detail is visible on the image. Typically, objects from a given class may have distinctive traits (e.g. a particular tail shape that allows to differentiate between two species of birds), but those can be more or less visible in the pictures. When those traits are present, the annotators can confidently predict the right class. Otherwise, they will make errors towards adjacent classes. In this case, the probability that the assigned label is wrong highly depends the instance (with distinctive traits being visible or not). Nonetheless, conditionally on the instance being corrupted, i.e., because those traits were not visible enough on the image, the transition probabilities to the adjacent classes are not influenced by the instance itself.

With the previous assumption, we obtain $\forall i \neq j, \forall x \in \mathcal{X}, T_{i,j}(x) = \alpha_{i,j}(1 - T_{i,i}(x))$ with $\alpha_{i,j} \in [0, 1]$. Once the $K(K-1)$ constants $(\alpha_{i,j})_{i \neq j}$ are estimated, we can derive the non-diagonal noise functions of $T(x)$ directly from our estimates of the diagonal noise functions (Eq. (5)).

3.2. Instance-level forward correction algorithm

Estimating $T_{i,i}$ and β_i. To train a classifier h with the *instance-level forward correction* method, we need to estimate both $T_{i,i}(x)$ and $\beta_i(x) = \frac{P(\bar{Y}=i|X=x)}{P(Y=i|X=x)}$ from Eq. (3),

for all $x \in S_i$. Firstly, the noisy posterior $P(\bar{Y} = i|X = x)$ can be easily estimated by training a naive classifier on the noisy dataset. Secondly, the true posterior $P(Y = i|X = x)$ can be estimated using the output of the classifier $h(x) = \hat{P}(Y = i|X = x)$ at the previous epoch.

Therefore, we iteratively update $\hat{\beta}$ and $\hat{T}$ with the following steps:

1. $\forall x \in \mathcal{X}$, initialize $\hat{\beta}_i(x) = 1$ and train a naive classifier h_{noisy} on the noisy data $\bar{D}$ to obtain $h_{\text{noisy}}(x) = \hat{P}(\bar{Y}|X = x)$.

2. $\forall i \in [1, K]$, for each sample $(x, \bar{y}, r_x) \in S_i$, compute $\hat{T}_{i,i}(x) = r_x \hat{\beta}_i(x)$ and train classifier h for one epoch.

3. $\forall i \in [1, K]$, for each sample $(x, \bar{y}, r_x) \in S_i$, update $\hat{\beta}_i(x) = \frac{h_{\text{noisy}}(x)_i}{h(x)_i}$.

Then, we repeat steps 2) and 3) through training. In this way, for every epoch, each function $T_{i,i}(\cdot)$ is estimated for the samples from S_i. Lastly, for the rest of samples with noisy label $j \neq i$, $T_{i,i}(\cdot)$ is estimated at each epoch using Eq. (4): $\forall (x, y, r_x) \in S \backslash S_i$,

$$\hat{T}_{i,i}(x) = \frac{1}{|S_i|} \sum_{(x', \bar{y}', r_x') \in S_i} r_x' \hat{\beta}_i(x') = \mu_i, \quad (6)$$

that is, the empirical mean of the values of $T_{i,i}(\cdot)$ on the subset of samples with noisy label i. Here we use this simple heuristic in order to keep our method straightforward but one could think of other methods to improve this estimation, such as using only the nearest neighbours of each sample or weighting the average by a measure of similarity between x and x'.

Computing $\alpha_{i,j}$. The computation of $\alpha_{i,j}$ boils down to approximating non-diagonal terms of the transition matrix in the CCN model. As

$$\forall i \neq j, \forall x \in \mathcal{X}, \; T_{i,j}(x) = \alpha_{i,j}(1 - T_{i,i}(x)),$$

we have:

$$\mathbb{E}_x[T_{i,j}(x)] = \alpha_{i,j}(1 - \mathbb{E}_x[T_{i,i}(x)])$$
$$\Leftrightarrow \alpha_{i,j} = \frac{\mathbb{E}_x[T_{i,j}(x)]}{1 - \mathbb{E}_x[T_{i,i}(x)]}.$$

A simple and reliable way is to use anchor points, i.e., points for which we can know the true class almost surely. These points may be directly available when some training data has been curated, or they can be identified either theoretically as in (Liu & Tao, 2015) or heuristically as in (Patrini et al., 2017). Having $S_i^* := \{(x, \bar{y}, r_x) \in S | P(Y = i|X = x) \approx$

1} a set of class i anchor points, we simply need compute: $\forall (x, \bar{y}, r_x) \in S_i^*, \forall j \neq i,$

$$\begin{cases} T_{i,i}(x) = r_x P(\bar{Y} = i|X = x) \\ T_{i,j}(x) = P(\bar{Y} = j|X = x) \end{cases}.$$

Two noisy posteriors can be estimated using the same classifier h_{noisy} trained on the noisy distribution $h_{\text{noisy}}(x) = \hat{P}(\bar{Y}|X = x)$ aforementioned. Thus, $\alpha_{i,j}$ can be estimated as follows: $\forall 1 \leq i, j \leq K, j \neq i,$

$$\alpha_{i,j} = \frac{\frac{1}{|S_i^*|} \sum_{(x, \bar{y}, r_x) \in S_i^*} h_{\text{noisy}}(x)_j}{1 - \frac{1}{|S_i^*|} \sum_{(x, \bar{y}, r_x) \in S_i^*} r_x h_{\text{noisy}}(x)_i}. \quad (7)$$

Summary of the training procedure. Given samples S and K sets of anchor points $(S_i^*)_{i=1}^K$, we want to train a classifier $h(\cdot)$ equipped with a loss l. For any loss $l : y, \hat{y} \mapsto l(y, \hat{y})$, we define the T-corrected loss as $l_T : y, \hat{y} \mapsto l(y, T\hat{y})$. The overall procedure is summarized in Algorithm 1.

Algorithm 1 Instance-Level Forward Correction (ILFC).

Input confidence-annotated samples $S := \{(x_i, \bar{y}_i, r_{x_i}), i = 1, \dots, N\}$, any loss l, classifier $h(\cdot)$, and anchor points sets $(S_i^*)_{i=1}^K$;
(1) Train a naive classifier h_{noisy} on samples $\{(x_i, \bar{y}_i)\}_{i=1}^N$;
(2) $\forall 1 \leq i, j \leq K, i \neq j$, compute $\alpha[i, j]$ from Eq. (7) with anchor points set S_i^*;
(3) $\forall 1 \leq i \leq K$, initialize $\beta_i(\cdot) = 1$;
for *epoch* $N = 1, \dots, N_{\max}$ **do**
    // Update diagonal constants
    (4) $\forall 1 \leq i \leq K$, compute $\mu[i]$ from Eq. (6);
    for $(x, \bar{y}, r_x) \in S$ **do**
        Set $i = \bar{y}$;
        // Compute diagonal terms
        (5) Set $T[i, i] = r_x \beta_i(x)$ and $\forall k \neq i, T[k, k] = \mu[k]$;
        // Compute non-diagonal terms
        (6) Set $\forall i, j \in \{1, \dots, K\}$, s.t. $i \neq j$, $T[i, j] = \alpha[i, j](1 - T[i, i])$;
        // Train classifier with
            instance-level corrected loss
        (7) Train $h(\cdot)$ on sample $(x, \bar{y}, r_x)$ with loss l_T;
        // Update density ratio estimate
        (8) Update $\forall 1 \leq i \leq K, \forall x \in S_i, \beta_i(x) = \frac{h_{\text{noisy}_i}(x)}{h_i(x)}$;
    end
end
(9) Output classifier $h(\cdot)$.

---

4. Experiments

We compare our instance-level forward correction (ILFC) method with four representative baselines: *forward correction* (FC) (Patrini et al., 2017), *mean absolute error* (MAE) (Ghosh et al., 2017), L_q-*norm* (LQ) (Zhang

& Sabuncu, 2018) and *co-teaching* (CT) (Han et al., 2018b). Details are shown in Appendix C. Note that the pioneer IDN methods cannot work for multi-class cases. The code is available at `https://github.com/antoninbrthn/CSIDN`.

4.1. Empirical Results on Synthetic Dataset

Generation process. We generate a synthetic dataset (Appendix B) consisting in three classes of concentric circles (Figure 7a). We then apply the following instance-dependent noise to each label: $P(\bar{Y} \neq Y|X = x) = \rho \left(\frac{w \cdot x}{\|w\|\|x\|} + 1 \right)/2$ with $w = (0, 1)$ and ρ controlling the mean noise rate. If corrupted, each label is flipped to another class uniformly.

Empirical results. Figure 3 shows the test accuracy of different methods on the synthetic dataset. Each experiment is repeated 5 times and we plot the confidence intervals of each curve. On low-level noise, all methods show good performances (Figure 3b). On mild-level noise, both Co-teaching and ILFC show good performances and outperform other baselines (Figure 3c). On high-level noise, the performance of all the baselines collapse, whereas ILFC constantly maintains good performances (Figures 3d and 3e).

Sensitivity analysis. In practice, the confidence scores obtained may not be accurate. Therefore, we conduct a sensitivity analysis to evaluate the robustness of ILFC: Similarly to Ishida et al. (2018), we add a zero-mean Gaussian noise with standard deviation $\sigma \in \{0.0, 0.3, 0.6\}$ to each confidence score and clip the values between 0 and 1. The bottom row of Figure 3 provides the resulting performances on the synthetic dataset, where ILFC shows good robustness to inaccurate confidence scores even with high standard deviation on a highly noisy dataset.

Limitation analysis. To explore and understand the limitation of different methods, we visualize their decision boundaries under synthetic IDN noise with high-level noise rate. Figure 4 shows the learned decision boundaries of our approach vs. the ones of a benchmark method (i.e., L_q norm), under 40% and 50% of IDN noise. With high-level noise rates, LQ method, which does not include any instance-level modelling, will degenerate around the most noisy region (i.e., upper region) of the input space. By contrast, our method successfully accounts for the high noise issue in upper region, keeping consistent predictions.

4.2. Empirical Results on Real Datasets

Generation process. In order to corrupt labels from clean datasets such as SVHN and CIFAR10, we adopt the following procedure: (1) train a classifier $h : x \mapsto \sigma(g(x))$ on a small subset of the clean dataset; (2) using a small validation set, calibrate the classifier by selecting the temperature t that maximizes the expected calibration error; (3) for each instance x, set: $\bar{y} = \text{argmax}_i\ h_t(x)_i$ and $r_x = \max_i\ h_t(x)_i$. With this process, we attempt to emulate the construction of a real-world dataset (Appendix D).

Main results analysis. Figures 5b and 5c show the test accuracy on SVHN with 25% and 45% instance-dependent noise, respectively. We can clearly observe that, on both low-level and high-level noise, ILFC shows good performances with a fast convergence rate, and outperforms other baselines. Figures 5d and 5e show the test accuracy on CIFAR10 with 25% and 45% instance-dependent noise, respectively. On low-level noise, all methods show good performances. However, on high-level noise, ILFC shows a fast convergence rate and outperforms other baselines.

Performance on clean data. Additionally, we run ILFC on the original noise-free datasets and compare it to a simple CCE loss (see Figure 6). Both methods demonstrate very similar performance in the noise-free case.

4.3. Empirical Results on Clothing1M

Finally, we demonstrate the effectiveness of our method on Clothing1M (Tong Xiao et al., 2015), an established real-world dataset in the label-noise learning literature. This dataset includes one million images of shopping items scrapped on the web and labeled automatically using surrounding meta-data. Because of the way this dataset is constructed, it involves instant-dependent label noise at its core: more ambiguous and ill-described items are more inclined to receive wrong labels. Additionally, curated labels are provided for a subset of around $50K$ samples.

Experimental method. We attach confidence scores to every instances of Clothing1M's noisy training set using the following steps: (1) fine-tune a naive Resnet-18 model pretrained on Imagenet with Clothing1M's clean training set; (2) calibrate this naive classifier using the clean validation set; (3) assign a confidence score to each instance of the noisy training set by using the naive classifier's softmax probability corresponding to the noisy label.

We use Resnet-18 models pretrained on ImageNet for both the main classifier and the naive classifier h_{noisy} and apply random crops of size 224×224 and horizontal flips as data augmentation. We use Adam with parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$ as the optimizer during training. The learning rate is set to 1.0×10^{-4} and decreased by a factor of 10 every 15 epochs. We set the batch size to 64. The benchmark methods are reimplemented and trained with a Resnet-18 for comparison. All models are trained on a NVIDIA Tesla K80 GPU. Additionally, we compare with the performance
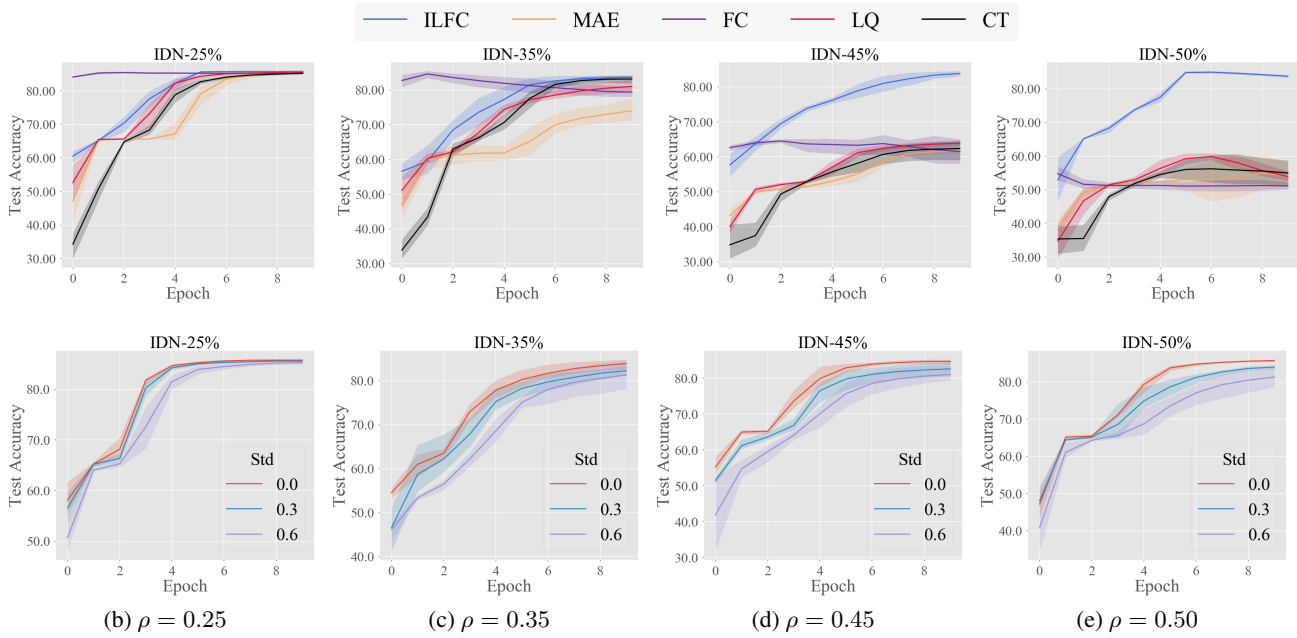
Figure 3. Empirical results on synthetic dataset. Top: The test accuracy on synthetic datasets with different levels of IDN noise. Bottom: Sensitivity analysis on the synthetic dataset: a zero-mean Gaussian noise of standard deviation $\sigma \in \{0.0, 0.3, 0.6\}$ is added to each confidence score before running ILFC with the noisy confidence scores.
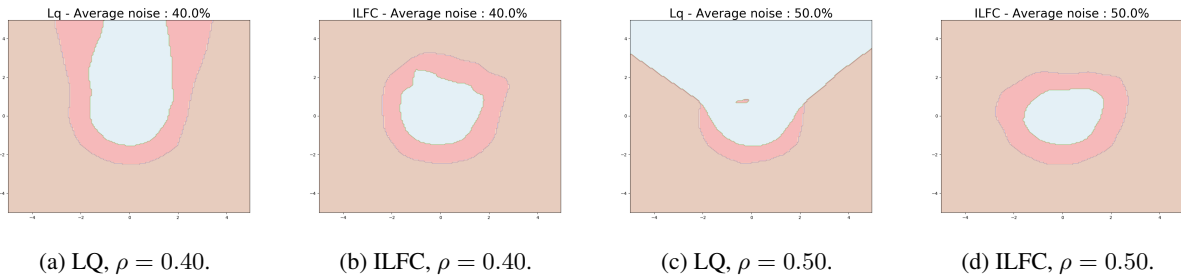

(a) LQ, $\rho = 0.40$.   (b) ILFC, $\rho = 0.40$.   (c) LQ, $\rho = 0.50$.   (d) ILFC, $\rho = 0.50$.

Figure 4. Decision boundaries of the learned classifier via our ILFC method and the L_q norm method (i.e., LQ). Under high-level noise rates (i.e., 40% and 50%), the LQ method degenerates around the most noisy region (i.e., upper region) of the input space, since it does not model any instance-level information essentially. However, our ILFC approach successfully addresses the high noise issue in upper region, staying consistent with the clean distribution.
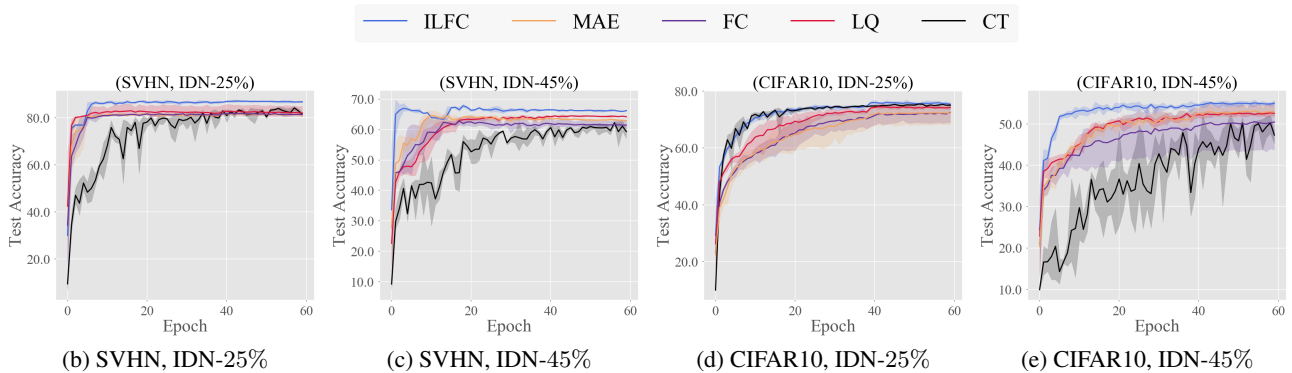

(b) SVHN, IDN-25%   (c) SVHN, IDN-45%   (d) CIFAR10, IDN-25%   (e) CIFAR10, IDN-45%

Figure 5. The test accuracy on real-world datasets with different levels of IDN noise.

Table 2. Test accuracy on Clothing1M. JoCor performance is taken from (Wei et al., 2020).

Method	CCE	Forward	MAE	LQ	Co-teaching	JoCor	ILFC
Accuracy	60.18	60.62	60.02	67.65	70.11	70.30	**73.35**

(a) CIFAR10   (b) SVHN
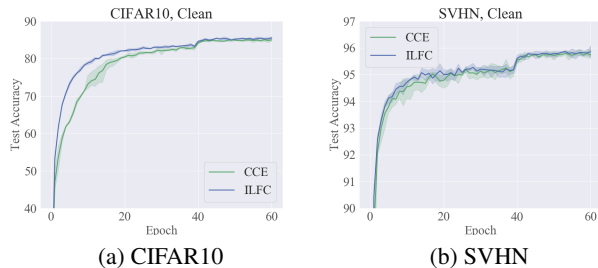
Figure 6. Test accuracy on noise-free datasets.

of the JoCor algorithm reported in (Wei et al., 2020), which is also based on a ResNet-18 architecture.

Results. As shown in Table 2, ILFC clearly outperforms benchmark methods. This shows the applicability of our method even to dataset not equipped with confidence scores in the first place: when a subset of curated data is available, it is possible to derive confidence scores from a calibrated classifier trained on a clean subset of the data. ILFC can then run successfully using those estimated confidence scores.

5. Conclusion

In this paper, we give an overview of label-noise learning from class-conditional noise (easier) to instance-dependent noise (harder). We explain why existing approaches cannot handle instance-dependent noise well, and try to address this challenge via confidence scores. Thus, we formally propose the _confidence-scored instance-dependent noise_ (CSIDN) model. To tackle the CSIDN model, we design a practical algorithm termed _instance-level forward correction_ (ILFC). Our ILFC method robustly outperforms existing methods, especially in the case of high-level noise and unknown real-world noise, even on datasets which are not equipped with confidence scores to begin with. In future works, we plan to extend label correction and sample selection approaches with the confidence scores from the CSIDN model.

Acknowledgements

References

Angluin, D. and Laird, P. Learning from noisy examples. *Machine Learning*, 2(4):343–370, 1988.

Arachie, C. and Huang, B. Adversarial label learning. In *AAAI*, 2019.

Arpit, D., Jastrzebski, 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. In *ICML*, 2017.

Bartlett, P. L., Jordan, M. I., and McAuliffe, J. D. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

Bootkrajang, J. and Chaijaruwanich, J. Towards instance-dependent label noise-tolerant classification: a probabilistic approach. *Pattern Analysis and Applications*, pp. 1–17, 2018.

Branson, S., Van Horn, G., and Perona, P. Lean crowdsourcing: Combining humans and machines in an online system. In *CVPR*, 2017.

Charoenphakdee, N., Lee, J., and Sugiyama, M. On Symmetric Losses for Learning from Corrupted Labels. *ICML*, 2019.

Cheng, H., Zhu, Z., Li, X., Gong, Y., Sun, X., and Liu, Y. Learning with instance-dependent label noise: A sample sieve approach. 2021.

Cheng, J., Liu, T., Ramamohanarao, K., and Tao, D. Learning with bounded instance and label-dependent label noise. In *ICML*, 2020.

Du, J. and Cai, Z. Modelling class noise with symmetric and asymmetric distributions. In *AAAI*, 2015.

Ghosh, A., Manwani, N., and S. Sastry, P. Making risk minimization tolerant to label noise. *Neurocomputing*, 160, 2014.

Ghosh, A., Kumar, H., and Sastry, P. Robust loss functions under label noise for deep neural networks. In *AAAI*, 2017.

Gneiting, T. and Raftery, A. E. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

Goldberger, J. and Ben-Reuven, E. Training deep neural-networks using a noise adaptation layer. In *ICLR*, 2017.

Guan, M. Y., Gulshan, V., Dai, A. M., and Hinton, G. E. Who said what: Modeling individual labelers improves classification. In *AAAI*, 2018.

Han, B., Yao, J., Niu, G., Zhou, M., Tsang, I., Zhang, Y., and Sugiyama, M. Masking: A new perspective of noisy supervision. In *NeurIPS*, 2018a.

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

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 *ICML*, pp. 4006–4016. PMLR, 2020.

Ishida, T., Niu, G., and Sugiyama, M. Binary classification from positive-confidence data. In *NeurIPS*, 2018.

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

Kearns, M. Efficient noise-tolerant learning from statistical queries. *STOC*, 1993.

Khetan, A., Lipton, Z. C., and Anandkumar, A. Learning from noisy singly-labeled data. In *ICLR*, 2018.

Krizhevsky, A. Learning multiple layers of features from tiny images. *Master's thesis, University of Tront*, 2009.

Laine, S. and Aila, T. Temporal Ensembling for Semi-Supervised Learning. *ICLR*, 2017.

Li, X., Liu, T., Han, B., Niu, G., and Sugiyama, M. Provably end-to-end label-noise learning without anchor points. In *ICML*, 2021.

Liu, T. and Tao, D. Classification with noisy labels by importance reweighting. *IEEE Transactions on pattern analysis and machine intelligence*, 38(3):447–461, 2015.

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.

Manwani, N. and Sastry, P. S. Noise tolerance under risk minimization. *IEEE Transactions on Cybernetics*, 43:1146–1151, 2013.

Masnadi-shirazi, H. and Vasconcelos, N. On the Design of Loss Functions for Classification: theory, robustness to outliers, and SavageBoost. In *NeurIPS*. 2009.

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

Menon, A. K., van Rooyen, B., and Natarajan, N. Learning from binary labels with instance-dependent noise. *Machine Learning*, 107(8-10):1561–1595, September 2018.

Miyato, T., Maeda, S.-i., Koyama, M., and Ishii, S. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993, 2018.

Natarajan, N., Dhillon, I. S., Ravikumar, P. K., and Tewari, A. Learning with Noisy Labels. In *NeurIPS*. 2013.

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., and Ng, A. Y. Reading digits in natural images with unsupervised feature learning. 2011.

Nock, R. and Nielsen, F. On the efficient minimization of classification calibrated surrogates. In *NeurIPS*, 2009.

Patrini, G., Nielsen, F., Nock, R., and Carioni, M. Loss factorization, weakly supervised learning and label noise robustness. In *ICML*, 2016.

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*, 2017.

Ratner, A., Bach, S. H., Ehrenberg, H., Fries, J., Wu, S., and Ré, C. Snorkel: Rapid training data creation with weak supervision. *The VLDB Journal*, 29(2):709–730, 2020.

Ratner, A. J., De Sa, C. M., Wu, S., Selsam, D., and Ré, C. Data programming: Creating large training sets, quickly. In *NeurIPS*, 2016.

Raykar, V. C., Yu, S., Zhao, L. H., Jerebko, A., Florin, C., Valadez, G. H., Bogoni, L., and Moy, L. Supervised learning from multiple experts: whom to trust when everyone lies a bit. In *ICML*, 2009.

Reed, S., Lee, H., Anguelov, D., Szegedy, C., Erhan, D., and Rabinovich, A. Training deep neural networks on noisy labels with bootstrapping. 2015.

Reid, M. D. and Williamson, R. C. Composite binary losses. *Journal of Machine Learning Research*, 11(Sep):2387–2422, 2010.

Scott, C. A generalized neyman-pearson criterion for optimal domain adaptation. In *ALT*, 2019.

Scott, C., Blanchard, G., and Handy, G. Classification with asymmetric label noise: Consistency and maximal denoising. In *COLT*, 2013.

Shen, Y. and Sanghavi, S. Learning with bad training data via iterative trimmed loss minimization. In *ICML*, 2019.

Shimodaira, H. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.

Snow, R., O'Connor, B., Jurafsky, D., and Ng, A. Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks. In *EMNLP*, 2008.

Stempfel, G. and Ralaivola, L. Learning SVMs from sloppily labeled data. In *International Conference on Artificial Neural Networks*, pp. 884–893, 2009.

Sugiyama, M. and Kawanabe, M. *Machine Learning in Non-Stationary Environments: Introduction to Covariate Shift Adaptation*. MIT Press, Cambridge, Massachusetts, USA, 2012.

Sugiyama, M., Krauledat, M., and Müller, K.-R. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(May):985–1005, 2007.

Sukhbaatar, S., Bruna, J., Paluri, M., Bourdev, L., and Fergus, R. Training convolutional networks with noisy labels. *ICLR workshop*, 2015.

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

Tarvainen, A. and Valpola, H. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS*, 2017.

Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. Learning from massive noisy labeled data for image classification. In *CVPR*, 2015.

van Rooyen, B. and Williamson, R. C. Learning in the Presence of Corruption. *Journal of Machine Learning Research*, 18, 2018.

Wang, Q., Han, B., Liu, T., Niu, G., Yang, J., and Gong, C. Tackling instance-dependent label noise via a universal probabilistic model. 2021.

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

Wu, S., Xia, X., Liu, T., Han, B., Gong, M., Wang, N., Liu, H., and Niu, G. Class2simi: A new perspective on learning with label noise. In *ICML*, 2021.

Xia, X., Liu, T., Han, B., Wang, N., Gong, M., Liu, H., Niu, G., Tao, D., and Sugiyama, M. Part-dependent label noise: Towards instance-dependent label noise. In *ICML*, volume 33, 2020.

Xia, X., Liu, T., Han, B., Gong, C., Wang, N., Ge, Z., and Chang, Y. Robust early-learning: Hindering the memorization of noisy labels. In *ICLR*, 2021.

Yan, Y., Rosales, R., Fung, G., Schmidt, M., Hermosillo, G., Bogoni, L., Moy, L., and Dy, J. Modeling annotator expertise: Learning when everybody knows a bit of something. In *AISTATS*, 2010.

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

Yao, Y., Liu, T., Han, B., Gong, M., Deng, J., Niu, G., and Sugiyama, M. Dual t: Reducing estimation error for transition matrix in label-noise learning. In *NeurIPS*, 2020b.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. In *ICLR*, 2017.

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

Zhang, T. et al. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32(1):56–85, 2004.

Zhang, Y., Niu, G., and Sugiyama, M. Learning noise transition matrix from only noisy labels via total variation regularization. In *ICML*, 2021.

Zhang, Z. and Sabuncu, M. Generalized Cross Entropy Loss for Training Deep Neural Networks with Noisy Labels. In *NeurIPS*. 2018.

Zhu, Z., Liu, T., and Liu, Y. A second-order approach to learning with instance-dependent label noise. In *CVPR*, 2021.