Mitigating Label Noise through Data Ambiguation

Julian Lienen¹, Eyke Hüllermeier²,³

¹Department of Computer Science, Paderborn University
²Institute of Informatics, LMU Munich
³Munich Center for Machine Learning

julian.lienen@upb.de, eyke@lmu.de

Abstract

Label noise poses an important challenge in machine learning, especially in deep learning, in which large models with high expressive power dominate the field. Models of that kind are prone to memorizing incorrect labels, thereby harming generalization performance. Many methods have been proposed to address this problem, including robust loss functions and more complex label correction approaches. Robust loss functions are appealing due to their simplicity, but typically lack flexibility, while label correction usually adds substantial complexity to the training setup. In this paper, we suggest to address the shortcomings of both methodologies by “ambiguated” the target information, adding additional, complementary candidate labels in case the learner is not sufficiently convinced of the observed training label. More precisely, we leverage the framework of so-called superset learning to construct set-valued targets based on a confidence threshold, which deliver imprecise yet more reliable beliefs about the ground-truth, effectively helping the learner to suppress the memorization effect. In an extensive empirical evaluation, our method demonstrates favorable learning behavior on synthetic and real-world tasks, confirming the effectiveness in detecting and correcting erroneous training labels.

Introduction

Label noise refers to the presence of incorrect or unreliable annotations in the training data, which can negatively impact the generalization performance of learning methods. Dealing with label noise constitutes a major challenge for the application of machine learning methods to real-world applications, which often exhibit noisy annotations in the form of erroneous labels or distorted sensor values. This is no less a concern for large models in the regime of deep learning, which have become increasingly popular due to their high expressive power, and which are not immune to the harming nature of label noise either. Existing methods addressing this issue include off-the-shelf robust loss functions, e.g., as proposed by Wang et al. (2019), and label correction, for example by means of replacing labels assumed to be corrupted (Wu et al. 2021). While the former appeals with its effortless integration into classical supervised learning setups, methods of the latter kind allow for a more effective suppression of noise by remodeling the labels. However, this comes at the cost of an increased model complexity, typically reducing the efficiency of the training (Liu et al. 2020).

The training dynamics of models have been thoroughly studied in the aforementioned regime (Chang, Learned-Miller, and McCallum 2017; Arazo et al. 2019; Liu et al. 2020), and two learning phases have been identified: Initially, as shown in the left plot of Fig. 1, the model shows reasonable learning behavior by establishing correct relations between features and targets, classifying even mislabeled instances mostly correctly. In this phase, the loss minimization is dominated by the clean fraction, such that the mislabeling does not affect the learning too much. However, after the clean labels are fit sufficiently well, the learner starts to concentrate predominantly on the mislabeled part, thereby overfitting incorrect labels and harming generalization.

A closer look at the probabilistic predictions in the first learning phase reveals that erroneous training labels can be distinguished from the underlying ground-truth classes by the learner’s confidence. The right plot in Fig. 1 illustrates that models are typically not incentivized to optimize for predicting the corrupted training labels in the first epochs, but infer relatively high probability scores for the (unknown) ground-truth class, at least on average. Evidently, the model

Figure 1: For ResNet34 models trained with cross-entropy on CIFAR-10 with 25% of corrupted instances (averaged over five seeds), the left plot shows the fractions of examples that are correctly classified, whose corrupted training label are memorized, or incorrectly classified with a label other than the ground-truth or training label, confirming the result in (Liu et al. 2020). The right plot illustrates the predicted probability magnitudes for clean or noisy labels.
itself could serve as a source for modeling the beliefs about the ground-truth, taking all (mostly clean) instance-label
relations into account to reason about the true labels.

This idea has been adopted by label correction methods that predict pseudo-labels for instances that appear to be
mislabeled, thus suggesting labels that the learner considers
more plausible (Reed et al. 2015; Wu et al. 2021). While intu-
tively plausible, a strategy like this is obviously not with-
out risk. Especially in the early phase of the training, replac-
ing a supposedly incorrect label by another one might be
too hasty, and may potentially even aggravate the negative
effects of label noise. Instead, as the learner is still in its in-
fancy, it seems advisable to realize a more cautious learning
behavior. More concretely, instead of discarding the original
training information completely, it might be better to keep it
as an option.

Following this motivation, we propose a complementary
approach for modeling the learner’s belief about the true la-
bel: Instead of forcing it to commit to a single label, either
the original one or a presumably more plausible alternative,
we allow the learner to (re-)label a training instance by a
set of candidate labels, which may comprise more than a
single plausible candidate. More specifically, by retaining
the original label and adding other plausible candidates, we
deliberately “ambiguous” the training information to miti-
gate the risk of making a mistake in an early phase of the
training process. To put this idea into practice, we make
use of so-called superset learning (Liu and Dietterich 2014;
Hüllermeier and Cheng 2015), which allows the learner it-
self to “disambiguate” possibly ambiguous training data.

More precisely, we represent the ambiguous target informa-
tion in the form of so-called credal sets, i.e., sets of prob-
ability distributions, to train probabilistic classifiers via gen-
eralized risk minimization in the spirit of label relaxation
(Lienen and Hüllermeier 2021a). We realize our approach,
which we dub Robust Data Ambiguity (RDA), in an easy
off-the-shelf loss function that dynamically derives the tar-
get sets from the model predictions without the need of any
additional model parameter – this is implicitly done in the
loss calculation, without requiring any change to a conven-
tional learning routine. This way, we combine the simplicity
of robust losses with the data modeling capabilities of more
complex label correction approaches. We demonstrate the
effectiveness of our method on commonly used image classi-
fication datasets with both synthetic and real-world noise,
confirming the adequacy of our proposed robust loss.

### Related Work

Coping with label noise in machine learning is a broad field
with an extensive amount of recent literature. Here, we dis-
tinguish four views on this issue, namely, robust loss func-
tions, regularization, sample selection, and label correction
methods. For a more comprehensive overview, we refer to
recent surveys by Song et al. (2020) and Wei et al. (2022b).

#### Robust Losses.

The task of designing robust optimization criteria has a long-standing history in classical statistics, e.g., to alleviate the sensitivity towards outliers. As a prominent member of such methods, the mean absolute er-
ror (MAE) steps up to mitigate the shortcomings of the mean
squared error. When relating to the context of probabilis-
tic classification, robustness of loss functions towards label
noise is linked with the symmetry of the function (Ghosh,
Kumar, and Sastry 2017), leading to adaptations of losses
such as cross-entropy (Wang et al. 2019; Ma et al. 2020).
A large strain of research proposes to balance MAE and
cross-entropy, e.g., by the negative Box-Cox transforma-
tion (Zhang and Sabuncu 2018), or controlling the order of
Taylor series for the categorical cross-entropy (Feng et al.
2020), whereas also alternative loss formulations have been
considered (Wei and Liu 2021). Besides, methodologies ac-
companying classical losses for robustness have been pro-
posed, such as gradient-clipping (Menon et al. 2020) or sub-
gradient optimization (Ma and Fattahi 2022).

#### Regularization.

Regularizing losses for classification has also been consid-
ered as a means to cope with label noise. As one prominent example, label smoothing (LS) (Szegedy
et al. 2016) has shown similar beneficial properties as loss
correction when dealing with label noise (Luksik et al.
2020). Advancements also enable applicability in high noise
regimes (Wei et al. 2022a). Among the first works build-
ing upon this observation, Arpit et al. (2017) characterize
two phases in learning from data with label noise. First,
the model learns to correctly classify most of the instances
(including the ground-truth labels of misclassified train-
ing instances), followed by the memorization of mislabels.

The works shows that explicit regularization, for instance
Dropout (Srivastava et al. 2014), is effective in combating
memorization and improving generalization. Following this,
Liu et al. (2020) propose a penalty term to counteract mem-
orization that stems from the (more correct) early learning
of the model. Beyond LS, also other forms of “soft labeling
methods” have been proposed, e.g., as in so-called label dis-
tribution learning (Gao et al. 2017). In addition, sparse reg-
ularization enforces the model to predict sharp distributions
with sparsity (Zhou et al. 2021b). Lastly, Iscen et al. (2022)
describe a regularization term based on the consistency of
an instance with its neighborhood in feature space.

#### Sample Selection.

Designed to prevent the aforementioned memorization of mislabeled instances, a wide variety
of methods rely on the so-called small loss selection
criterion (Jiang et al. 2018; Gwi, Wang, and Tian 2021).
It is intended to distinguish clean samples the model recog-
nizes well, and from which it can learned without distor-
tion. This distinction suggests a range of approaches, in-
cluding a gradual increase of the clean sample set with in-
creasing model confidence (Shen and Sanghavi 2019; Cheng
et al. 2021; Wang et al. 2022), co-training (Han et al.
2018; Yu et al. 2019; Wei et al. 2020), or re-weighting in-
stances (Chen, Zhu, and Chen 2021). Furthermore, it makes
the complete plethora of classical semi-supervised learning
methods amenable to the task of learning from noisy labels.
Here, the non-small loss examples are considered as unla-
bled in the first place (Li, Socher, and Hoi 2020; Nishi et al.
2021; Zheltonozhskii et al. 2022). Often, such methodology
is also combined with consistency regularization (Bachman,
Alsharif, and Precup 2014), as prominently used in classical

**Label Correction.** Traditionally, correcting label noise has also been considered by learning probabilistic transition matrices (Goldberger and Ben-Reuven 2017; Patrini et al. 2017; Zhu, Wang, and Liu 2022; Kye et al. 2022), e.g., to re-weight or correct losses. In this regard, it has been proposed to learn a model and (re-)label based on the model’s predictions jointly (Tanaka et al. 2018). Closely related to sample selection as discussed before, Song, Kim, and Lee (2019) propose to first select clean samples in a co-teaching method, which then refurbishes noisy samples by correcting their label information. Furthermore, Arazo et al. (2019) suggest to use two-component mixture models for detecting mislabels that are to be corrected, thus serving as a clean sample criterion. Wu et al. (2021) and Tu et al. (2023) approach the problem of label correction by a meta-learning technique in the form of a two-stage optimization process of the model and the training labels. In addition, Liu et al. (2022) describe an approach to learn the individual instances’ noise by an additional over-parameterization. Finally, ensembling has also been considered as a source for label correction (Lu and He 2022).

**Method**

The following section outlines the motivation behind our method RDA to model targets in an ambiguous manner, introduces the theoretical foundation, and elaborates on how this approach realizes robustness against label noise.

**Motivation**

Deep neural network models, often overparameterized with significantly more parameters than required to fit training data points at hand, have become the de-facto standard choice for most practically relevant problems in deep learning. Here, we consider probabilistic models of the form \( \bar{p} : X \rightarrow \mathbb{P}(Y) \) with \( X \subset \mathbb{R}^d \) and \( Y = \{y_1, \ldots, y_K\} \) being the feature and target space, respectively, and \( \mathbb{P}(Y) \) the class of probability distributions on \( Y \). We denote the training set of size \( N \) as \( \mathcal{D}_N = \{(x_i, y_i)\}_{i=1}^N \subset X \times Y \), where each instance \( x_i \in X \) is associated with an underlying (true) label \( y_i \). The latter is conceived as being produced by the underlying (stochastic) data-generating process, i.e., as the realization of a random variable \( Y \sim \bar{p}(\cdot | x) \).\(^1\) However, the actual label \( y_i \) encountered in the training data might be corrupted, which means that \( y_i \neq y_i^* \). Reasons for corruption can be manifold, including erroneous measurement and annotation by a human labeler.

As real-world data typically comprises noise in the annotation process, e.g., due to human labeling errors, their robustness towards label noise is of paramount importance. Alas, models in the regime of deep learning trained with conventional (probabilistic) losses, such as cross-entropy, lack this robustness. As thoroughly discussed in the analysis of the training dynamics of such models (Chang, Learned-Miller, and McCallum 2017; Arazo et al. 2019; Liu et al. 2020), two phases can be distinguished in the training of alike models, namely a “correct concept learning” and a *memorization* (Arpit et al. 2017) or forgetting (Toneva et al. 2019) phase (cf. Fig. 1). While the model relates instances to labels in a meaningful way in the first phase, leading to correct predictions on most training instances, the transition to a more and more memorizing model harms the generalization performance.

Looking closer at the learning dynamics in an idealized setting, one can observe that overparameterized models project instances of the same ground-truth class \( y^* \) to a similar feature embedding, regardless of having observed a correct or corrupted training label in the first training phase. Fig. 2 illustrates the learned feature activations of the penultimate layer of a multi-layer perceptron with a classification head trained on MNIST over the course of the training (see the appendix for further experimental details). In the beginning, the learner predicts a relatively high probability \( \tilde{p}(y_i^* | x_i) \) even for noisy instances \( (x_i, y_i) \), as the loss is dominated by the cross-entropy of the clean instances, which leads to a similar marginalization behavior as in the non-corrupted case. Here, the proximity of mislabeled instances to the decision boundary can be related to the “hardness” of instances to be learned, as also studied by Chang, Learned-Miller, and McCallum (2017). In later stages, the feature activations of the noisy instances shift in a rotating manner, successively being pushed towards the other discriminatory face of the decision boundary. This goes hand in hand with a decrease and eventual increase of the predicted probability scores \( \max_{y \in Y} \tilde{p}(y | x) \), consistent with the observations described in the appendix.

---

\(^1\)In applications such as image classification, the ground-truth conditional distributions \( p^*(\cdot | x) \in \mathbb{P}(Y) \) are often degenerate, i.e., \( p^*(y^* | x) \approx 1 \) and \( p^*(y | x) \approx 0 \) for \( y \neq y^* \).
made in Fig. 1.

It appears natural to seek for means that keep corrupted instances \((\cdot, y_i)\) close to clean samples \((\cdot, y'_i)\) in the feature space, and not let the learning bias in the later stages pull them over towards the label corruption. In the context of the overall optimization, the model itself represents a source of justification for considering a class label as a plausible outcome. Hence, we suggest to use the model predictions simultaneously with the training labels for modeling our beliefs about the ground-truth. Consequently, we shall consult not only the individual training labels, but also the complete training dataset that found its way into the current model in conjunction, complementing the former as a second piece of evidence. This represents a distillation of knowledge obtained from the (mostly clean) data at hand, and we argue that the confidence \(\hat{p}\) is a suitable surrogate to represent plausibility in this regard.

Credal Labeling

Inquiring the model prediction \(\hat{p}(x_i)\) in addition to the training label \(y_i\) may call for augmenting the (hit)herto single) label by an additional plausible candidate label \(\hat{y}_i \in \arg\max_{y' \in \mathcal{Y}} \hat{p}(y' | x_i)\) with \(\hat{y}_i \neq y_i\), making the target effectively ambiguous and hence less convenient for conventional point-wise classification losses. However, from a data modeling perspective, it is important to recognize that the imprecisiation of the target information pays off with higher validation, i.e., it is more likely that the true label is covered by the target labels. This is completely in line with the idea of data imprecisiation in the context of so-called superset learning (Lienen and Hüllemeier 2021b). Thus, ambiguating presumably corrupt training information appears to be a useful means to counter the influence of mislabeling.

Before detailing the representation of the aforementioned beliefs, we shall revisit the conventional probabilistic learning setting. To train probabilistic classifiers \(\hat{p}\) as specified above, e.g., by minimizing the cross-entropy loss, traditional methods transform deterministic target labels \(y_i \in \mathcal{Y}\) as commonly observed in classification datasets into degenerate probability distributions \(p_{y_i} \in \mathcal{P}(\mathcal{Y})\) with \(p_{y_i}(y_i) = 1\) and \(p_{y_i}(y) = 0\) for \(y \neq y_i\). As a result, the predicted distribution \(\hat{p}\) can be compared to \(p_{y_i}\) using a probabilistic loss \(\mathcal{L} : \mathcal{P}(\mathcal{Y}) \times \mathcal{P}(\mathcal{Y}) \rightarrow \mathbb{R}\) to be minimized. It is easy to see that \(p_{y_i}\) assigns full plausibility to the observed training label \(y_i\), while the other labels are regarded as fully implausible.

Ambiguity in a probabilistic sense can be attained by enriching the representation of the target distribution in a set-valued way. To this end, Lienen and Hüllemeier (2021a) propose to use credal sets, i.e., sets of probability distributions, as a means to express beliefs about the true class conditional distribution \(p^*\). This allows one, for example, to not only consider \(p_{y_i}\) for the training label \(y_i\) as a fully plausible target distribution, but also \(p_{\hat{y}_i}\) for a plausible candidate label \(\hat{y}_i \neq y_i\) and interpolations between these extremes, i.e., distributions \(p = \lambda p_{y_i} + (1 - \lambda)p_{\hat{y}_i}\) assigning probability \(\lambda\) to label \(y_i\) and \(1 - \lambda\) to \(\hat{y}_i\). More generally, a credal set can be any (convex) set of probability distributions, including, for example, “relaxations” of extreme distributions \(p_{y_i}\), which reduce the probability for \(y_i\) to \(1 - \epsilon < 1\) and reassign the remaining mass of \(\epsilon > 0\) to the other labels. Such distributions appear to be meaningful candidates as target distributions, because a degenerate ground truth \(p^* = p_{y_i}\) is actually not very likely (Lienen and Hüllemeier 2021a).

Possibility theory (Dubois and Prade 2004) offers a compact representation of credal sets in the form of upper bounds on probabilities, encoded as possibility distributions \(\pi : \mathcal{Y} \rightarrow [0, 1]\). Intuitively, for each outcome \(y \in \mathcal{Y}\), \(\pi(y)\) defines an upper bound on the true probability \(p^*(y)\). More generally, a distribution \(\pi\) induces a possibility measure \(\Pi(Y) = \max_{y \in Y} \pi(y)\) on \(\mathcal{Y}\), which in turn defines an upper bound on any event \(Y \subseteq \mathcal{Y}\). In our case, a distribution \(\pi_i\) associated with an instance \(x_i\), assigns a possibility (or plausibility) to each class in \(\mathcal{Y}\) for being the true outcome associated with \(x_i\), and \(\pi_i(y)\) can be interpreted as an upper bound on \(p^*(y | x_i)\). Thus, the credal set of distributions considered as candidates for \(p^*(y | x_i)\) is given as follows:

\[
Q_{\pi_i} := \left\{ p \in \mathcal{P}(\mathcal{Y}) | \forall Y \subseteq \mathcal{Y} : \sum_{y \in Y} p(y) \leq \max_{y \in Y} \pi_i(y) \right\}
\]

(1)

Data Ambiguation for Robust Learning

With the above approach to modeling ambiguous targets in a probabilistic manner, we can put the idea of a robust ambiguization method on a rigorous theoretic foundation. As a concrete model of the target label for a training instance \((x_i, y_i)\), we propose the following confidence-thresholded possibility distribution \(\pi_i\):

\[
\pi_i(y) = \begin{cases} 
1 & \text{if } y = y_i \lor \hat{p}(y | x_i) \geq \beta \\
\alpha & \text{otherwise}
\end{cases}
\]

(2)

where \(\beta \in [0, 1]\) denotes the confidence threshold for the prediction \(\hat{p}(y | x_i)\), and \(\alpha \in [0, 1)\) is the label relaxation parameter (Lienen and Hüllemeier 2021a). Thus, while the original label \(y_i\) remains fully plausible (\(\pi_i(y_i) = 1\)), it might be complemented by other candidates as soon as the predicted probability for another label exceeds \(\beta\), it is deemed a fully plausible candidate, too. Besides, by assigning a small degree of possibility \(\alpha\) to all remaining classes, these are not completely excluded either. Both \(\alpha\) and \(\beta\) are treated as hyperparameters.

To learn from credal targets \(Q_{\pi_i}\) induced by \(\pi_i\), i.e., from the ambiguaged training data \(\{(x_i, Q_{\pi_i})\}\), we make use of a method for superset learning that is based on minimizing the optimistic superset loss (Hüllemeier 2014). The latter generalizes a standard loss \(\mathcal{L}\) for “point predictions” as follows:

\[
\mathcal{L}^*(Q_\pi, \hat{p}) := \min_{p \in Q_\pi} \mathcal{L}(p, \hat{p}).
\]

(3)

As argued by Hüllemeier (2014), minimizing OSL on the ambiguaged training data is suitable for inducing a model while simultaneously disambiguating the data. In other words, the learner makes a most favourable selection among the candidate labels, and simultaneously fits a model to that selection (by minimizing the original loss \(\mathcal{L}\)).

A common choice for \(\mathcal{L}\) is the Kullback-Leibler divergence \(D_{KL}\), for which (3) simplifies to

\[
\mathcal{L}^*(Q_\pi, \hat{p}) = \begin{cases} 
0 & \text{if } \hat{p} \in Q_{\pi} \\
D_{KL}(p^* || \hat{p}) & \text{otherwise}
\end{cases}
\]

(4)
Figure 3: A barycentric visualization of the confidence-thresholded ambiguity for a corrupt training label \( y_1 \) and a ground-truth \( y_2 \) in the target space \( \mathcal{Y} = \{ y_1, y_2, y_3 \} \). Starting from a credal set \( \mathcal{Q} \) centered at \( p_{y_1} \) (left plot), the prediction \( \tilde{p} \) predicts a probability mass greater than \( \beta \) for \( y_2 \). Consequently, full possibility is assigned to \( y_2 \), leading to \( \mathcal{Q} \) as shown to the right.

**Algorithm 1 Robust Data Ambiguation (RDA) Loss**

**Require:** Training instance \((x, y) \in \mathcal{X} \times \mathcal{Y}, \) model prediction \( \tilde{p}(x) \in \mathcal{P}(\mathcal{Y}), \) confidence threshold \( \beta \in [0, 1] \), relaxation parameter \( \alpha \in (0, 1] \)

1. Construct \( \pi \) as in Eq. (4) with
   \[
   \pi(y') = \begin{cases} 
   1 & \text{if } y' = y \lor \tilde{p}(y' | x) \geq \beta \\
   \alpha & \text{otherwise}
   \end{cases}
   \]

2. return \( \mathcal{L}^*(Q_\pi, \tilde{p}(x)) \) as specified in Eq. (4), where \( Q_\pi \) is derived from \( \pi \)

where
   \[
   p^\pi(y) = \begin{cases} 
   (1 - \alpha) \cdot \frac{\tilde{p}(y)}{\sum_{y' \in \mathcal{Y}, \pi(y') = 1} \tilde{p}(y')}) & \text{if } \pi_i(y) = 1 \\
   \alpha \cdot \frac{\sum_{y' \in \mathcal{Y}, \pi_i(y') = \alpha} \tilde{p}(y')}{\sum_{y' \in \mathcal{Y}, \pi_i(y') = \alpha} p(y')} & \text{otherwise}
   \end{cases}
   \]

is projecting \( \tilde{p} \) onto the boundary of \( Q_\pi \). This loss is provably convex and has the same computational complexity as standard losses such as cross-entropy (Lien and Hüllermeier 2021a).

Fig. 3 illustrates the core idea of our method, which we will refer to as **Robust Data Ambiguation (RDA).** Imagine an instance \( x \) with true label \( y_2 \) but corrupted training information \( y_1 \). Initially, we observe a probabilistic target centered at \( p_{y_1} \) (and potentially relaxed by some \( \alpha > 0 \)). Without changing the target, label relaxation would compute the loss by comparing the prediction \( \tilde{p} \) to a (precise) distribution \( p^\pi \) projecting \( \tilde{p} \) towards \( p_{y_1} \). In this example, the model predicts \( y_2 \) with a confidence exceeding \( \beta \). With our method, full plausibility would be assigned to \( y_2 \) in addition to \( y_1 \), leading to a credal set \( \mathcal{Q} \) as shown in the right plot. To compute the loss, \( \tilde{p} \) is now compared to a less extreme target \( p^\pi \) that is projected onto the larger credal set, no longer urging the learner to predict distributions close to \( p_{y_1} \). A more technical description of RDA is provided in Alg. 1, confirming the simplicity of our proposal.

As can be seen, by ambiguating the target set, we relieve the learner from memorizing wrong training labels. At the same time, we remain cautious and try to avoid committing to potentially erroneous predictions. Thereby, learning from noisy labels is de-emphasized while the optimization on clean samples remains unaffected. At the same time, driven by the loss minimization on similar instances that are correctly labeled with \( y_2 \), the imprecisation allows the predictions \( \tilde{p} \in \mathcal{Q} \) to evolve towards the true target \( p_{y_2} \).

Surpassing the parameter \( \beta \) can be seen as an indication for mislabeling, i.e., the prediction \( \tilde{p} \) suggests that a different label \( y \neq y_1 \) is a plausible candidate for representing the true outcome \( y^* \). As such, \( \beta \) can be used to control the “cautiousness” in suspecting mislabeling: High values for \( \beta \) force that only highly confident predictions shall adjust the target, whereas less extreme values result in a more eager addition of candidates. Conversely, since classes \( y \) with \( \tilde{p}(y) < \beta \) are considered as implausible, the threshold \( \beta \) could also be interpreted as a criterion for exclusion of candidates. As shown in a more extensive analysis of the \( \beta \) parameter, the former interpretation is practically more useful in the case of robust classification.

Generally, any value in \([0, 1]\) could be chosen for \( \beta \), perhaps also depending on the training progress. As a simple yet effective rule of thumb, we suggest to vary \( \beta \) over the course of time: As the model is relatively uncertain in the first epochs, one should not spent too much attention to the predictions. However, with further progress, \( \beta \) becomes more informative for the confidence-driven possibility elicitation, suggesting to use smaller \( \beta \) values. Empirically, we found the following decay function to yield good performance:

\[
\beta_T = \beta_1 + \frac{1}{2} (\beta_0 - \beta_1) \left(1 + \cos \left(\frac{T}{T_{\text{max}}} \pi\right)\right),
\]

where \( T \) and \( T_{\text{max}} \) denote the current and maximum number of epochs, respectively, while \( \beta_0 \) and \( \beta_1 \) represent the start and end values for \( \beta \). Nevertheless, as will be shown in the empirical evaluation, also static values for \( \beta \) work reasonably well, such that the number of additional hyperparameters can be reduced to a single one.

**Experiments**

To demonstrate our method’s effectiveness in coping with label noise, we conduct an empirical analysis on several image classification datasets as a practically relevant problem domain, although RDA is not specifically tailored to this domain and does not use any modality-specific components. Here, we consider CIFAR-10/-100 (Krizhevsky and Hinton 2009) as well as the large-scale datasets WebVision (Li et al. 2017) and Clothing1M (Xiao et al. 2015) as benchmarks.

For the former, we model synthetic noise by both symmetrically and asymmetrically randomizing the labels for a fraction of examples, while WebVision and Clothing1M comprise real-world noise by their underlying crawling process. Moreover, we report results on CIFAR-10/0(N) (Wei et al. 2022b) as another real-world noise dataset based on human annotators. We refer to the appendix for a more detailed description of the datasets and the corruption process, as well as experiments on additional benchmark datasets.

As baselines, we take a wide range of commonly applied loss functions into account. Proceeding from the conven-
tional cross-entropy (CE), we report results for the regularized CE adaptations label smoothing (LS) and label relaxation (LR), as well as the popular robust loss functions generalized cross-entropy (GCE) (Zhang and Sabuncu 2018), normalized cross-entropy (NCE) (Ma et al. 2020), combinations of NCE with AUL and AGCE (Zhou et al. 2021a) and CORES (Cheng et al. 2021). For completeness, we also report results for ELR (Liu et al. 2020) and SOP (Liu et al. 2022) as two state-of-the-art representatives of regularization and label correction methods, respectively. While all other losses can be used off-the-shelf, these two have additional parameters (to track the label noise per instance), giving them an arguably unfair advantage. In the appendix, we show further results for a natural baseline to our approach in the realm of superset learning, as well as experiments that combine our method with sample selection.

We follow common practice and evaluate methods against label noise in the regime of overparameterized models by training ResNet34 models on the smaller scale datasets CIFAR-10/-100. For the larger datasets, we consider ResNet50 models pretrained on ImageNet. All models use the same training procedure and optimizer. A more thorough overview of experimental details, such as hyperparameters and their optimization, as well as the technical infrastructure, can be found in the appendix. We repeated each run five times with different random seeds, reporting the accuracy on the test splits to quantify generalization.

**Synthetic Noise**

Table 1 reports the results for the synthetic corruptions on CIFAR-10/-100. As can be seen, our approach provides consistent improvements in terms of generalization performance over the robust off-the-shelf loss functions. Cross-entropy and its regularized adaptations LS and LR appear sensitive to label noise, which confirms the need for loss robustness. Interestingly, although being slightly inferior in most cases of symmetric noise, our method appears still competitive compared to ELR and SOP, despite their increased expressivity through additional parameters. For asymmetric noise, as presented in the appendix, our method could even outperform such methods. Nevertheless, it becomes less effective with an increasing level of noise.

When looking at the learning dynamics of training with RDA, Fig. 4 reveals an effective ambiguation in the course of the learning process. The left plot shows the attenuation of any memorization while improving the correctness of model predictions for the wrongly labeled instances at the same time. The plot in the middle depicts an increase of the credal set size in terms of classes with full plausibility for the noisy instances. Together with the right plot showing the validity of the credal sets, i.e., the fraction of credal sets that assign full possibility to the ground-truth class, one can easily see that the ambiguation is indeed able to select the ground-truth class as training label. Furthermore, the credal set size for clean instances is barely affected, which again confirms the adequacy of our model. Notably, our method also shows self-correcting behavior after ambiguating with a wrong class midway. While the validity of the credal sets increases (roughly) monotonically, the set sizes become smaller towards the end of the training. In the appendix, we provide additional plots in other noise settings showing consistent effects.

**Real-World Noise**

In coherence with the previous observations, our robust ambiguation loss also works reasonably well for real-world noise. As presented in Table 2, RDA leads to superior generalization performance compared to baseline losses without any additional parameters in almost any case. Moreover, it also consistently outperforms SOP on CIFAR-10N, whereas it leads to similar results for CIFAR-100N.

For the large-scale datasets WebVision and Clothing1M, for which results are presented in Table 3, the differences between the baselines and our approach appear to be rather modest but are still in favor of our method.

**Conclusion**

Large models are typically prone to memorizing noisy labels in classification tasks. To address this issue, various loss functions have been proposed that enhance the robustness of conventional loss functions against label noise. Although such techniques are appealing due to their simplicity, they typically lack the capacity to incorporate additional knowledge regarding the instances, such as beliefs about the true label. In response, pseudo-labeling methods for label correction have emerged. They allow for modeling target information in a more sophisticated way, albeit at the cost of an
Table 1: Test accuracies and standard deviations on the test split for models trained on CIFAR-10(0) with (symmetric) synthetic noise. The results are averaged over runs with different seeds, bold entries mark the best method without any additional model parameters. Underlined results indicate the best method overall.

Table 2: Test accuracies and standard deviations on the test split for models trained on CIFAR-10(0)N with real-world noise, the scores on the clean splits can be found in the appendix. The results are averaged over runs with different seeds.

Table 3: Large-scale test accuracies on WebVision and Clothing1M using ResNet50 models. The baseline results for the two datasets are taken from (Zhou et al. 2021a) and (Liu et al. 2020), respectively.

increased training complexity.

To address the shortcomings of previous methods, our approach advocates the idea of weakening (ambiguating) the training information. In a sense, it unifies the two directions pursued so far, namely, modifying the learning procedure (loss function) and modifying the data (selection, correction). By allowing the learner to (re-)label training data with credal sets of probabilistic labels, the approach becomes very flexible. For the specific type of label ambiguity we proposed in this paper, we could show that learning from the re-labeled data is equivalent to minimizing a "robustified" loss function that is easy to compute. Our empirical evaluation confirms the adequacy of our proposal.

Our approach suggests several directions for future research. For example, a more informed choice of $\beta$ and $\alpha$ could be realized by quantifying the epistemic uncertainty (Hüllermeier and Waegeman 2021) of individual predictions, as highly uncertain guesses should be considered with caution. Also, our method could be leveraged for various downstream tasks, e.g., to detect anomalies in mostly homogeneous data. Finally, as we are currently only considering the predictions at a time, further improvements could be achieved by taking the dynamics of the training into account.
Acknowledgements

This work was partially supported by the German Research Foundation (DFG) within the Collaborative Research Center “On-The-Fly Computing” (CRC 901 project no. 160364472). Moreover, the authors gratefully acknowledge the funding of this project by computing time provided by the Paderborn Center for Parallel Computing (PC²).

References


Chang, H.; Learned-Miller, E. G.; and McCallum, A. 2017. Active Bias: Training More Accurate Neural Networks by Emphasizing High Variance Samples. In NeurIPS, Dec. 4-9, Long Beach, CA, USA, 1002–1012.


