Peer Loss Functions: Learning from Noisy Labels without Knowing Noise Rates

Yang Liu [1] **Hongyi Guo** [2]

Abstract

Learning with noisy labels is a common challenge in supervised learning. Existing approaches often require practitioners to specify *noise rates*, i.e., a set of parameters controlling the severity of label noises in the problem, and the specifications are either assumed to be given or estimated using additional steps. In this work, we introduce a new family of loss functions that we name as *peer loss functions*, which enables learning from noisy labels and does not require a priori specification of the noise rates. Peer loss functions work within the standard empirical risk minimization (ERM) framework. We show that, under mild conditions, performing ERM with peer loss functions on the noisy data leads to the optimal or a near-optimal classifier as if performing ERM over the clean training data, which we do not have access to. We pair our results with an extensive set of experiments. Peer loss provides a way to simplify model development when facing potentially noisy training labels, and can be promoted as a robust candidate loss function in such situations.

1. Introduction

The quality of supervised learning models depends on the quality of the training dataset $\{(x_n, y_n)\}_{n=1}^N$. In practice, label noise can arise due to a host of reasons. For instance, the observed labels $\tilde{y}_n$s may represent human observations of a ground truth label. In this case, human annotators may observe the label imperfectly due to differing degrees of expertise or measurement error, see e.g., medical examples such as labeling MRI images from patients. There exist extensive prior works in the literature that aim to develop algorithms to learn models that are robust to label noise (Bylander, 1994; Cesa-Bianchi et al., 1999; 2011; Ben-David et al.; Scott et al., 2013; Natarajan et al., 2013; Scott, 2015). Typical solutions that have theoretical guarantees often require *a priori* knowledge of *noise rates*, i.e., a set of parameters that control the severity of label noise. Working with unknown noise rates is difficult in practice: Usually, one must estimate the noise rates from data, which may require additional data collection or requirement (Natarajan et al., 2013; Scott, 2015; Van Rooyen et al., 2015a) (e.g., a set of ground truth labels for tuning these parameters) and may introduce estimation error that can affect the final model in less predictable ways. Our main goal is to provide an alternative that does not require the specification of the noise rates, nor an additional estimation step for the noise. This target solution benefits the practitioner when he or she does not have access to reliable estimates of the noise rates (e.g., when the training data has a limited size for the estimation tasks, or when the training data is already collected in a form that makes the estimation hard to perform).

In this paper, we introduce a new family of loss functions, *peer loss functions*, to empirical risk minimization (ERM), for a broad class of learning with noisy labels problems. Peer loss functions operate under different noise rates without requiring either *a priori* knowledge of the embedded noise rates, or an estimation procedure. This family of loss functions builds on approaches developed in the *peer prediction* literature (Miller et al., 2005; Dasgupta & Ghosh, 2013; Shnayder et al., 2016), which studies how to elicit information from self-interested agents without verification. Results in the peer prediction literature focused on designing scoring functions to score each reported data using another noisy reference answer, without accessing ground truth information. We borrow this idea and the associated scoring functions via making a connection through treating each classifier's predictions as an agent's private information to be elicited and evaluated, and the noisy labels as imperfect reference answers reported from a "noisy label agent". The specific form of peer loss evaluates classifiers' prediction using noisy labels on both the samples to-be-evaluated and carefully constructed "peer" samples. The evaluation on the constructed peer sample encodes *implic-*

[1]Computer Science and Engineering, UC Santa Cruz, Santa Cruz, CA, USA [2]Computer Science and Engineering, Shanghai Jiao Tong University, China. Correspondence to: Yang Liu <yangliu@ucsc.edu>, Hongyi Guo <guohongyi@sjtu.edu.cn>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

itly the information about the noise as well as the underlying true labels, which helps us offset the effects of label noise. The peer sample evaluation returns us a favorable property that the expected risk of peer loss computed on the noisy distribution turns to be an affine transformation of the true risk of the classifier defined on the clean distribution. In other words, peer loss is invariant to label noise when optimizing with it. This effect helps us get rid of the estimation of noise rates.

The main contributions of this work are:

1. We propose a new family of loss functions that can easily adapt to existing ERM framework that i) is robust to *asymmetric* label noise with formal theoretical guarantees and ii) requires no prior knowledge or estimation of the noise rates (*no need for specifying noise rates*). We believe having the second feature above is non-trivial progress, and it features a promising solution to deploy in an unknown noisy training environment.
2. We present formal results showing that performing ERM with a peer loss function can recover an optimal, or a near-optimal classifier f^* as if performing ERM on the clean data (Theorem 2, 3, 4). We also provide peer loss functions' risk guarantees (Theorem 5, 7).
3. We present extensive experimental results to validate the usefulness of peer loss functions (Section 5 and Appendix). This result is encouraging as it demonstrates the practical effectiveness in removing the requirement of error rates of noise before many of the existing training methods can be applied. We also provide preliminary results on how peer loss generalizes to multi-class classification problems.
4. Our implementation of peer loss functions is available at https://github.com/gohsyi/PeerLoss.

Due to space limit, the full version of this paper with all proof and experiment details can be found in (Liu & Guo, 2020).

1.1. Related Work

We go through the most relevant works.[1]

Learning from Noisy Labels Our work fits within a stream of research on learning with noisy labels. A large portion of research on this topic works with the *random classification noise* (RCN) model, where observed labels are flipped independently with probability $\in [0, \frac{1}{2}]$ (Bylander, 1994; Cesa-Bianchi et al., 1999; 2011; Ben-David et al.). Recently, learning with asymmetric noisy data (or also referred as *class-conditional* random classification noise (CCN)) for binary classification problems has been rigorously studied in (Stempfel & Ralaivola, 2009; Scott et al., 2013; Natarajan et al., 2013; Scott, 2015; Van Rooyen et al., 2015a; Menon et al., 2015).

For RCN, where the noise parameters are symmetric, there exist works that show symmetric loss functions (Manwani & Sastry, 2013; Ghosh et al., 2015; 2017; Van Rooyen et al., 2015a) are robust to the underlying noise, without specifying the noise rates. Our focus departs from this line of works and we exclusively focus on asymmetric noise setting, and study the possibility of an approach that can ignore the knowledge of noise rates. Follow-up works include (Du Plessis et al., 2013; Van Rooyen et al., 2015b; Menon et al., 2015; Charoenphakdee et al., 2019).

More Recent Works More recent developments include an importance re-weighting algorithm (Liu & Tao, 2016), a noisy deep neural network learning setting (Sukhbaatar & Fergus, 2014; Han et al., 2018; Song et al., 2019), and learning from massive noisy data for image classification (Xiao et al., 2015; Goldberger & Ben-Reuven, 2016; Zhang et al., 2017; Jiang et al., 2017; Jenni & Favaro, 2018; Yi & Wu, 2019), robust cross entropy loss for neural network (Zhang & Sabuncu, 2018), loss correction (Patrini et al., 2017), among many others. Loss or sample correction has also been studied in the context of learning with unlabeled data with weak supervisions (Lu et al., 2018). Most of the above works either lacks theoretical guarantees of the proposed method against asymmetric noise rates (Sukhbaatar & Fergus, 2014; Zhang & Sabuncu, 2018), or require estimating the noise rate (or transition matrix between the noisy and true labels) (Liu & Tao, 2016; Xiao et al., 2015; Patrini et al., 2017; Lu et al., 2018).

A recent work (Xu et al., 2019) proposes an information theoretical loss, an idea adapted from an earlier theoretical contribution (Kong & Schoenebeck, 2018), which is also robust to asymmetric noise rates. We aimed for a simple-to-optimize loss function that can easily adapt to existing ERM solutions.

Peer Prediction Our work builds on the literature of peer prediction (Prelec, 2004; Miller et al., 2005; Witkowski & Parkes, 2012; Radanovic & Faltings, 2013; Witkowski et al., 2013; Dasgupta & Ghosh, 2013; Shnayder et al., 2016; Liu & Chen, 2017). Most relevant to us is (Dasgupta & Ghosh, 2013; Shnayder et al., 2016) where a correlated agreement (CA) type of mechanism was proposed. CA evaluates a report's correlations with another reference agent - its specific form inspired our peer loss.

2. Preliminaries

Suppose $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ are drawn from a joint distribution $\mathcal{D}$, with their marginal distributions denoted as

[1] We provide more detailed discussions in the Appendix.

$\mathbb{P}_X, \mathbb{P}_Y$. We assume $\mathcal{X} \subseteq \mathbb{R}^d$, and $\mathcal{Y} = \{-1, +1\}$, that is we consider a binary classification problem. Denote by $p := \mathbb{P}(Y = +1) \in (0, 1)$. There are N training samples $(x_1, y_1), ..., (x_N, y_N)$ drawn i.i.d. from $\mathcal{D}$. For positive integer n, denote by $[n] := \{1, 2, ..., n\}$.

Instead of observing y_ns, the learner can only collect a noisy set of training labels $\tilde{y}_n$s, generated according to y_ns and a certain error rate model; that is we observe a dataset $\{(x_n, \tilde{y}_n)\}_{n=1}^N$. We assume uniform error for all the training samples we collect, in that errors in $\tilde{y}_n$s follow the same error rate model: denoting the random variable for noisy labels as $\tilde{Y}$ and we define

$$e_{+1} := \mathbb{P}(\tilde{Y} = -1|Y = +1),\ e_{-1} := \mathbb{P}(\tilde{Y} = +1|Y = -1)$$

Label noise is conditionally independent from the features, that is the error rate is uniform across x_ns: $\mathbb{P}(\tilde{Y} = y'|Y = y) = \mathbb{P}(\tilde{Y} = y'|X, Y = y), \forall y, y' \in \{-1, +1\}$.

We assume $0 \leq e_{+1} + e_{-1} < 1$ - this condition is not unlike the ones imposed in the existing learning literature (Natarajan et al., 2013), and it simply implies that the noisy labels are positively correlating with the true labels (informative about the true labels). Denote the distribution of the noisy data $(X, \tilde{Y})$ as $\tilde{\mathcal{D}}$.

$f : \mathcal{X} \to \mathbb{R}$ is a real-valued decision function, and its risk w.r.t. the 0-1 loss is defined as $\mathbb{E}_{(X,Y)\sim\mathcal{D}}[\mathbb{1}(f(X), Y)]$. The Bayes optimal classifier f^* is the one that minimizes the 0-1 risk: $f^* = \text{argmin}_f\, \mathbb{E}_{(X,Y)\sim\mathcal{D}}[\mathbb{1}(f(X), Y)]$. Denote this optimal risk as R^*. Instead of minimizing the above 0-1 risk, the learner often seeks a surrogate loss function $\ell : \mathbb{R} \times \{-1, +1\} \to \mathbb{R}_+$, and finds a $f \in \mathcal{F}$ that minimizes the following error: $\mathbb{E}_{(X,Y)\sim\mathcal{D}}[\ell(f(X), Y)]$. $\mathcal{F}$ is the hypothesis space for f. Denote the following measures: $R_{\mathcal{D}}(f) = \mathbb{E}_{(X,Y)\sim\mathcal{D}}[\mathbb{1}(f(X), Y)]$ and $R_{\ell,\mathcal{D}}(f) = \mathbb{E}_{(X,Y)\sim\mathcal{D}}[\ell(f(X), Y)]$.

When there is no confusion, we will also short-hand $\mathbb{E}_{(X,Y)\sim\mathcal{D}}[\ell(f(X), Y)]$ as $\mathbb{E}_{\mathcal{D}}[\ell(f(X), Y)]$. Denoting D a dataset collected from distribution $\mathcal{D}$ (correspondingly $\tilde{D} := \{(x_n, \tilde{y}_n)\}_{n\in[N]}$ from $\tilde{\mathcal{D}}$), the empirical risk measure for f is defined as $\hat{R}_{\ell,D}(f) = \frac{1}{|D|}\sum_{(x,y)\in D} \ell(f(x), y)$.

2.1. Learning with Noisy Labels

Typical methods for learning with noisy labels include developing noise correction surrogates loss function to learn with noisy data (Natarajan et al., 2013). For instance, (Natarajan et al., 2013) tackles this problem by defining the following *un-biased surrogate loss functions* over ℓ to help "remove" noise in expectation: $\tilde{\ell}(t, y) := \frac{(1-e_{-y})\cdot\ell(t,y) - e_y\cdot\ell(t,-y)}{1-e_{-1}-e_{+1}}, \forall t, y$. $\tilde{\ell}$ is identified such that when a prediction is evaluated against a noisy label using this surrogate loss function, the prediction is as if evaluated against the ground-truth label using ℓ in expectation. Hence the loss of the prediction is "unbiased", that is $\forall$ prediction t, $\mathbb{E}_{\tilde{Y}|y}[\tilde{\ell}(t, \tilde{Y})] = \ell(t, y)$ [Lemma 1, (Natarajan et al., 2013)].

One important note to make is most, if not all, existing solutions require the knowledge of the error rates e_{-1}, e_{+1}. Previous works either assumed the knowledge of it, or needed additional assumptions, clean labels or redundant noisy labels to estimate them. This becomes the bottleneck of applying these great techniques in practice. Our work is also motivated by the desire to remove this limitation.

2.2. Peer Prediction

Peer prediction is a technique developed to truthfully elicit information when there is no ground truth verification. Suppose we are interested in eliciting private observations about a binary event $Y \in \{-1, +1\}$ generated according to a random variable Y. There are K agents indexed by $[K]$. Each of them holds a noisy observation of the truth $Y = y$, denoted as $y^A \in \{-1, +1\}$, $A \in [K]$. We would like to elicit the y^As, but they are completely private and we will not observe y to evaluate agents' reports. Denote by r^A the reported data from each agent A. $r^A \neq y^A$ if agents are not compensated properly for their information.

Results in *peer prediction* have proposed scoring or reward functions that evaluate an agent's report using the reports of other peer agents. For example, a peer prediction mechanism may reward agent A for her report r^A using $S(r^A, r^B)$ where r^B is the report of a randomly selected reference agent $B \in [K]\backslash\{A\}$. The scoring function S is designed so that truth-telling is a strict Bayesian Nash Equilibrium (implying other agents truthfully report their y^B), that is, $\mathbb{E}_{y^B}[S(y^A, y^B)|y^A] > \mathbb{E}_{y^B}[S(r^A, y^B)|y^A]$, $\forall r^A \neq y^A$.

Correlated Agreement (Shnayder et al., 2016; Dasgupta & Ghosh, 2013) (CA) is an established peer prediction mechanism for a multi-task setting[2]. CA is also the core and the focus of our subsequent sections on developing peer loss functions. This mechanism builds on a Δ matrix that captures the stochastic correlation between the two sources of predictions y^A and y^B. Denote the following relabeling function: $g(1) = -1, g(2) = +1$, $\Delta \in \mathbb{R}^{2\times 2}$ is a squared matrix with its entries defined as follows: $\forall\, k, l = 1, 2$

$$\Delta_{k,l} = \mathbb{P}\big(y^A = g(k), y^B = g(l)\big) - \mathbb{P}\big(y^A = g(k)\big)\mathbb{P}\big(y^B = g(l)\big),$$

The intuition of above Δ matrix is that each (k, l) entry of Δ captures the marginal correlation between the two predictions y^A and y^B. When there is no confusion in the text, we will always follow this relabeling function to map a -1 label to 1 and $+1$ to 2 when defining or calling an entry in

[2] We provide other examples of peer prediction functions in the Appendix.

the Δ matrix without explicitly spelling out $g(\cdot)$, that is we will write

$$\Delta_{k,l} = \mathbb{P}\big(y^A = k, y^B = l\big) - \mathbb{P}\big(y^A = k\big) \cdot \mathbb{P}\big(y^B = l\big),$$

as well as

$$\Delta_{y,y'} = \mathbb{P}\big(y^A = y, y^B = y'\big) - \mathbb{P}\big(y^A = y\big) \cdot \mathbb{P}\big(y^B = y'\big).$$

We further define $M : \{-1, +1\} \times \{-1, +1\} \to \{0, 1\}$ as the sign matrix of Δ:

$$M(y, y') =: \text{Sgn}\left(\Delta_{y,y'}\right), \tag{1}$$

where $\text{Sgn}(x) = 1, x > 0$; $\text{Sgn}(x) = 0$, otherwise.

CA requires each agent A to perform multiple tasks: denote agent A's predictions for the N tasks as $y_1^A, \ldots, y_N^A$. Ultimately the scoring function $S(\cdot)$ for each task n that is shared between A, B is defined as follows: randomly draw two tasks n_1, n_2 , $n_1 \neq n_2$,

$$S\big(y_n^A, y_n^B\big) := M\big(y_n^A, y_n^B\big) - M\big(y_{n_1}^A, y_{n_2}^B\big).$$

A key difference between the first and second $M(\cdot)$ terms is that the second term is defined for two independent peer tasks n_1, n_2 (as the reference answers). It was established in (Shnayder et al., 2016) that CA is truthful at a Bayesian Nash Equilibrium (Theorem 5.2, (Shnayder et al., 2016).) [3]; in particular, if y^B is *categorical* w.r.t. y^A: $\mathbb{P}(y^B = y'|y^A = y) < \mathbb{P}(y^B = y'), \forall A, B \in [K]$, $y' \neq y$ then $S(\cdot)$ is strictly truthful (Theorem 4.4, (Shnayder et al., 2016)).

3. Learning with Noisy Labels: a Peer Prediction Approach

In this section, we show that peer prediction scoring functions, when specified properly, will adopt Bayes optimal classifier as their maximizers (or minimizers for the corresponding loss form).

3.1. Learning with Noisy Labels as an Elicitation Problem

We first state our problem of learning with noisy labels as a peer prediction problem. The connection is made by firstly rephrasing the two data sources, the classifiers' predictions and the noisy labels, from agents' perspective. For a task $Y \in \{-1, +1\}$, say $+1$ for example, denote the noisy labels $\tilde{Y}$ as $Z(X), X \sim \mathbb{P}_{X|Y=1}$. In general, $Z(X)$ can be interpreted as the agent that "observes" $\tilde{y}_1, ..., \tilde{y}_N$ for a set of randomly drawn feature vectors $x_1, ..., x_N$: $\tilde{y}_n \sim Z(X)$. Denote the following error rates for the agent's observations (similar to the definition of e_{+1}, e_{-}): $\mathbb{P}(Z(X) = -1|Y = +1) = e_{+1}$, $\mathbb{P}(Z(X) = +1|Y = -1) = e_{-1}$. There is another agent whose observations "mimic" the Bayes optimal classifier f^*. Again denote this optimal classifier agent as $Z^*(X) := f^*(X)$: $\mathbb{P}(Z^*(X) = -1|Y = +1) = e_{+1}^*$, $\mathbb{P}(Z^*(X) = +1|Y = -1) = e_{-1}^*$.

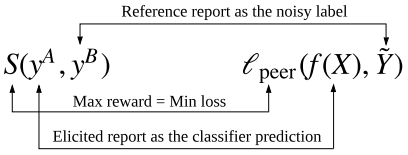

Figure 1. S is the peer prediction function; ℓ_{peer} is to "evaluate" a classifier's prediction using a noisy label.

Suppose we would like to elicit predictions from the optimal classifier agent Z^*, while the reports from the noisy label agent Z will serve as the reference reports. Both Z and Z^* are randomly assigned a task $X = x$, and each of them observes a signal $Z(x)$ and $Z^*(x)$ respectively. Denote the report from agent Z^* as r^*. A scoring function $S : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is called to induce truthfulness if the following fact holds: $\forall r^*(X) \neq Z^*(X)$,

$$\mathbb{E}_X\big[S\big(Z^*(X), Z(X)\big)\big] \geq \mathbb{E}_X\big[S\big(r^*(X), Z(X)\big)\big]. \tag{2}$$

Taking the negative of $S(\cdot)$ (changing a reward score one aims to maximize to a loss to minimize) we also have

$$\mathbb{E}_X\big[-S\big(Z^*(X), Z(X)\big)\big] \leq \mathbb{E}_X\big[-S\big(r^*(X), Z(X)\big)\big],$$

implying when taking $-S(\cdot)$ as the loss function, minimizing $-S(\cdot)$ w.r.t. Z will the Bayes optimal classifier f^*.

Our idea is summarized in Figure 1.

3.2. Peer Prediction Mechanisms Induce Bayes Optimal Classifier

When there is no ambiguity, we will shorthand $Z(X), Z^*(X)$ as Z, Z^*, with keeping in mind that Z, Z^* encode the randomness in X. In the elicitation setting, a potentially misreported classifier $f(X)$ only disagrees with $f^*(X)$ according to its local observation $f^*(X)$ but not Y (unobservable to the agent), that is $\mathbb{P}(f(X) \neq f^*(X)|f^*(X) = l, Y = +1) = \mathbb{P}(f(X) \neq f^*(X)|f^*(X) = l, Y = -1)$, $l \in \{-1, +1\}$. Denote this reporting space of f as $\mathcal{F}_{\text{report}}$: clearly f^* belongs to this space (truthful reporting). Suppose Z^* has the correct prior p of Y. Then we have:

Theorem 1. *Suppose $S(\cdot)$ induces truthful f^* (Eqn. (2)), that is $S(\cdot)$ is able to elicit the Bayes optimal classifier f^* (agent Z^*) using Z. Then $f^* = \text{argmin}_{f \in \mathcal{F}_{\text{report}}} \mathbb{E}_{(X,\tilde{Y}) \sim \tilde{\mathcal{D}}}\big[-S(f(X), \tilde{Y})\big]$.*

[3] To be precise, it is an informed truthfulness. We refer interested readers to (Shnayder et al., 2016) for details.

This proof can be done via showing that any non-optimal Bayes classifier corresponds to a non-truthful misreporting strategy. We emphasize that it is not super restrictive to have a truthful peer prediction scoring function S. We provide discussions in Appendix. Theorem 1 provides a conceptual connection and can serve as an anchor point when connecting a peer prediction score function to the problem of learning with noisy labels (with restriction to a particular reporting space). So far we have not discussed a specific form of how we construct a loss function using ideas from peer prediction, and have not mentioned the requirement of knowing the noise rates. We will provide the detail about a particular *peer loss* (with little to none restriction) in the next section, and explain its independence of noise rates.

4. Peer Loss Function

We now present peer loss, a family of loss functions inspired by a particular peer prediction mechanism, the correlated agreement (CA), as presented in Section 2.2. We are going to show that peer loss is able to induce the minimizer of a hypothesis space $\mathcal{F}$, under a broad set of non-restrictive conditions. In this Section, we do not restrict to Bayes optimal classifiers, nor do we impose any restrictions on the loss functions' elicitation power.

4.1. Preparing CA for Noisy Learning Problem

To give a gentle start, we repeat the setting of CA for our classification problem.

Δ and scoring matrix First recall that $\Delta \in \mathbb{R}^{2\times 2}$ is a squared matrix with entries defined between Z^* (the f^*) and Z (i.e., the noisy labels $\tilde{Y}$): $\forall k, l = 1, 2$

$$\Delta_{k,l} = \mathbb{P}\big(f^*(X) = k, \tilde{Y} = l\big) - \mathbb{P}\big(f^*(X) = k\big)\mathbb{P}\big(\tilde{Y} = l\big),$$

Δ characterizes the "marginal" correlations between the optimal classifier' prediction and the noisy label $\tilde{Y}$. Then the following scoring matrix M is computed using $\text{Sgn}(\Delta)$, the sign matrix of Δ.

Example 1. *Consider a binary class label case:* $\mathbb{P}(Y = -1) = 0.4, \mathbb{P}(Y = +1) = 0.6$*, the noise in the labels are* $e_{-1} = 0.3, e_{+1} = 0.4$ *and* $e^*_{-1} = 0.2, e^*_{+1} = 0.3$*. Then we have* $\Delta_{1,1} = 0.036$, $\Delta_{1,2} = -0.036$, $\Delta_{2,1} = -0.036$, $\Delta_{2,2} = 0.036$*. The details of the calculation can be found in the Appendix. And:* $\Delta = \begin{bmatrix} 0.036 & -0.036 \\ -0.036 & 0.036 \end{bmatrix} \Rightarrow Sgn(\Delta) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$.

Peer samples For each sample $(x_n, \tilde{y}_n)$, randomly draw another two samples $(x_{n_1}, \tilde{y}_{n_1}), (x_{n_2}, \tilde{y}_{n_2})$ such that $n_1 \neq n_2$. We will name $(x_{n_1}, \tilde{y}_{n_1}), (x_{n_2}, \tilde{y}_{n_2})$ as n's peer samples. After pairing x_{n_1} with $\tilde{y}_{n_2}$ (two independent instances), the scoring function $S(\cdot)$ for each sample point x_n is defined as follows:

$$S(f(x_n), \tilde{y}_n)) = M\big(f(x_n), \tilde{y}_n\big) - M\big(f(x_{n_1}), \tilde{y}_{n_2}\big).$$

Define loss function $\tilde{\ell}(\cdot)$ as the negative of $S(\cdot)$, which we will name as the (**Generic Peer Loss**)

$$\tilde{\ell}\big(f(x_n), \tilde{y}_n\big) := \big(1 - M\big(f(x_n), \tilde{y}_n\big)\big) - \big(1 - M\big(f(x_{n_1}), \tilde{y}_{n_2}\big)\big). \tag{3}$$

The first term above evaluates the classifier's prediction on x_n using noisy label $\tilde{y}_n$, and the second "peer" term defined on two independent tasks n_1, n_2 "punishes" the classifier from overly agreeing with the noisy labels. We will see this effect more clearly.

4.2. Peer Loss

We need to know $\text{Sgn}(\Delta)$ in order to specify M and $\tilde{\ell}$, which requires certain information about f^* and $\tilde{Y}$. We show that Example 1 is not a special case, and for the scenarios that the literature is broadly interested in, $\text{Sgn}(\Delta)$ is simply the identify matrix:

Lemma 1. *When* $e_{-1} + e_{+1} < 1$*, we have* $Sgn(\Delta) = I_{2\times 2}$*, the identity matrix.*

The above implies that for $\Delta_{k,k}, k = 1, 2$, f^* and $\tilde{Y}$ are positively correlated, so the marginal correlation is positive; while for off-diagonal entries, they are negatively correlated.

Peer Loss When $\text{Sgn}(\Delta) = I_{2\times 2}$, $M(y, y') = 1$ if $y = y'$, and 0 otherwise. $\tilde{\ell}(\cdot)$ defined in Eqn. (3) reduces to the following form:

$$\mathbb{1}_{\text{peer}}(f(x_n), \tilde{y}_n) = \mathbb{1}(f(x_n), \tilde{y}_n) - \mathbb{1}(f(x_{n_1}), \tilde{y}_{n_2}) \tag{4}$$

To see this, for instance $1 - M\big(f(x_n) = +1, \tilde{y}_n = +1\big) = 1 - M(2, 2) = 1 - 1 = 0 = \mathbb{1}(f(x_n) = +1, \tilde{y}_n = +1)$. Replacing $\mathbb{1}(\cdot)$ with any generic loss $\ell(\cdot)$ we define:

$$\ell_{\text{peer}}(f(x_n), \tilde{y}_n) = \ell(f(x_n), \tilde{y}_n) - \ell(f(x_{n_1}), \tilde{y}_{n_2}) \tag{5}$$

We name the above loss as *peer loss*. This strikingly simple form of $\ell_{\text{peer}}(f(x_n), \tilde{y}_n)$ implies that knowing $e_{-1} + e_{+1} < 1$ holds is all we need to specify ℓ_{peer}.

Later we will show this particular form of loss is invariant under label noise, which gives peer loss the ability to drop the requirement noise rates. We will instantiate this argument formally with Lemma 2 and establish a link between the above measure and the true risk of a classifier on the clean distribution. The rest of presentation focuses on ℓ_{peer} (Eqn. (5)), but ℓ_{peer} recovers $\mathbb{1}_{\text{peer}}$ via replacing ℓ with $\mathbb{1}$.

ERM with Peer Loss Performing ERM with peer loss returns us $\hat{f}^*_{\ell_{\text{peer}}}$:

$$\hat{f}^*_{\ell_{\text{peer}}} = \arg\min_{f\in\mathcal{F}} \frac{1}{N}\sum_{n=1}^{N} \ell_{\text{peer}}(f(x_n), \tilde{y}_n) \tag{6}$$

Note again that the definition of ℓ_{peer} does not require the knowledge of either e_{+1}, e_{-1} or e^*_{+1}, e^*_{-1}.

4.3. Property of Peer Loss

We now present a key property of peer loss, which shows that its risk over the noisy labels is simply an affine transformation of its true risk on clean data. We denote by $\mathbb{E}_{\mathcal{D}}[\ell_{\text{peer}}(f(X), Y)]$ the expected peer loss of f when (X, Y), as well as its peer samples, are drawn i.i.d. from distribution $\mathcal{D}$.

Lemma 2. *Peer loss is invariant to label noise:*

$$\mathbb{E}_{\tilde{\mathcal{D}}}[\ell_{peer}(f(X), \tilde{Y})] = (1-e_{-1}-e_{+1})\cdot\mathbb{E}_{\mathcal{D}}[\ell_{peer}(f(X), Y)].$$

The above Lemma states that peer loss is invariant to label noise in expectation. We have also empirically observed this effect in our experiment. Therefore minimizing it over noisy labels is equivalent to minimizing over the true clean distribution. The theorems below establish the connection between $\mathbb{E}_{\mathcal{D}}[\ell_{\text{peer}}(f(X), Y)]$, the expected peer loss over clean data, with the true risk:

Denote $\tilde{f}^*_{\mathbb{1}_{\text{peer}}} = \arg\min_{f\in\mathcal{F}} R_{\mathbb{1}_{\text{peer}},\tilde{\mathcal{D}}}(f)$. With Lemma 2, we can easily prove the following:

Theorem 2. *[Optimality guarantee with equal prior] When $p = 0.5$, $\tilde{f}^*_{\mathbb{1}_{peer}} \in \arg\min_{f\in\mathcal{F}} R_{\mathcal{D}}(f)$.*

The above theorem states that for a class-balanced dataset with $p = 0.5$, peer loss induces the same minimizer as the one that minimizes the 0-1 loss on the clean data. Removing the constraint of $\mathcal{F}$, i.e., $\tilde{f}^*_{\mathbb{1}_{\text{peer}}} = \arg\min_f R_{\mathbb{1}_{\text{peer}},\tilde{\mathcal{D}}}(f) \Rightarrow \tilde{f}^*_{\mathbb{1}_{\text{peer}}} = f^*$. In practice we can balance the dataset s.t. $p \to 0.5$. When $p \neq 0.5$, denote $\delta_p = \mathbb{P}(Y = +1) - \mathbb{P}(Y = -1)$, we prove:

Theorem 3. *[Approximate optimality guarantee with unequal prior] When $p \neq 0.5$, $|R_{\mathcal{D}}(\tilde{f}^*_{\mathbb{1}_{peer}}) - \min_{f\in\mathcal{F}} R_{\mathcal{D}}(f)| \leq |\delta_p|$.*

When $|\delta_p|$ is small, i.e., p is closer to 0.5, this bound becomes tighter.

Multi-class extension Our results in this section are largely generalizable to the multi-class classification setting. Suppose we have K classes of labels, denoting as $\{1, 2, ..., K\}$. One can show that for many classes of noise matrices, the $M(\cdot)$ matrix is again an identify matrix. This above fact will help us reach the conclusion that minimizing peer loss leads to the same minimizer on the clean data. We provide experiment results for multi-class tasks in Section 5.

Why do we not need the knowledge of noise rates explicitly? Both of the terms $\mathbb{1}(f(x_n), \tilde{y}_n)$ and $\mathbb{1}(f(x_{n_1}), \tilde{y}_{n_2})$ encoded the knowledge of noise rates ***implicitly***. The carefully constructed form as presented in Eqn. (4) allows peer loss to be invariant against noise (Lemma 2, a property we will explain later). For a preview, for example if we take expectation of $\mathbb{1}_{\text{peer}}(f(x_n) = +1, \tilde{y}_n = +1)$ we will have

$$\begin{aligned}&\mathbb{E}\left[\mathbb{1}_{\text{peer}}(f(x_n) = +1, \tilde{y}_n = +1)\right]\\ &= \mathbb{P}(f(X) = +1, \tilde{Y} = +1) - \mathbb{P}(f(X) = +1)\mathbb{P}(\tilde{Y} = +1),\end{aligned}$$

the marginal correlation between f and $\tilde{Y}$, which is exactly capturing the entries of Δ defined between f and $\tilde{Y}$! The second term above is a product of marginals because of the independence of peer samples n_1, n_2. Using the constructed peer term is all we need to recover this information measure in expectation. In other words, both the joint and marginal product distribution terms encode the noise rate information in an implicit way.

4.4. α-weighted Peer Loss

We take a further look at the case with $p \neq 0.5$. Denote by $R_{+1}(f) = \mathbb{P}(f(X) = -1|Y = +1)$, $R_{-1}(f) = \mathbb{P}(f(X) = +1|Y = -1)$. It is easy to prove:

Lemma 3. *Minimizing $\mathbb{E}[\mathbb{1}_{peer}(f(X), \tilde{Y})]$ is equivalent to minimizing $R_{-1}(f) + R_{+1}(f)$.*

However, minimizing the true risk $R_{\mathcal{D}}(f)$ is equivalent to minimizing $p\cdot R_{+1}(f) + (1-p)\cdot R_{-1}(f)$, a weighted sum of $R_{+1}(f)$ and $R_{-1}(f)$. The above observation and the failure to reproduce the strong theoretical guarantee when $p \neq 0.5$ motivated us to study a α-weighted version of peer loss, to make peer loss robust to the case $p \neq 0.5$. We propose the following *α-weighted peer loss* via adding a weight $\alpha \geq 0$ to the second term, the peer term:

$$\ell_{\alpha\text{-peer}}\big(f(x_n), \tilde{y}_n\big) = \ell(f(x_n), \tilde{y}_n) - \alpha\cdot\ell(f(x_{n_1}), \tilde{y}_{n_2})$$

Denote $\mathbb{1}_{\alpha\text{-peer}}$ as $\ell_{\alpha\text{-peer}}$ when $\ell = \mathbb{1}$, $\tilde{f}^*_{\mathbb{1}_{\alpha\text{-peer}}} = \arg\min_{f\in\mathcal{F}} R_{\mathbb{1}_{\alpha\text{-peer}},\tilde{\mathcal{D}}}(f)$ as the optimal classifier under $\mathbb{1}_{\alpha\text{-peer}}$, and $\delta_{\tilde{p}} = \mathbb{P}(\tilde{Y} = +1) - \mathbb{P}(\tilde{Y} = -1)$. Then when $\delta_{\tilde{p}} \neq 0$ (when this condition does not hold, we can perturb the training data by downsampling one of the two classes according to the noisy labels.), we prove:

Theorem 4. *Let $\alpha = 1 - (1 - e_{-1} - e_{+1})\cdot\frac{\delta_p}{\delta_{\tilde{p}}}$. We have $\tilde{f}^*_{\mathbb{1}_{\alpha\text{-peer}}} \in \arg\min_{f\in\mathcal{F}} R_{\mathcal{D}}(f)$.*

Denote $\alpha^* := 1 - (1 - e_{-1} - e_{+1})\cdot\frac{\delta_p}{\delta_{\tilde{p}}}$. Several remarks follow: (1) When $p = 0.5$, $\delta_p = 0$, we have $\alpha^* = 1$,

i.e. we recover the earlier definition of ℓ_{peer}. (2) When $e_{-1} = e_{+1}$, $\alpha^* = 0$ (see Appendix for details), we recover the ℓ for the clean learning setting, which has been shown to be robust under symmetric noise rates (Manwani & Sastry, 2013; Van Rooyen et al., 2015a). (3) When the signs of $\mathbb{P}(Y=1) - \mathbb{P}(Y=-1)$ and $\mathbb{P}(\tilde{Y}=1) - \mathbb{P}(\tilde{Y}=-1)$ are the same, $\alpha^* < 1$. Otherwise, $\alpha^* > 1$. In other words, when the label noise changes the relative quantitative relationship of $\mathbb{P}(Y=1)$ and $\mathbb{P}(Y=-1)$, $\alpha^* > 1$ and vice versa. (4) Knowing α^* requires a certain knowledge of e_{+1}, e_{-1} when $p \neq 0.5$. Though we do not claim this knowledge, this result implies tuning α^* (using validation data) may improve the performance.

Theorem 2 and 4 and sample complexity theories imply that performing ERM with $\mathbb{1}_{\alpha^*\text{-peer}}$: $\hat{f}^*_{\mathbb{1}_{\alpha^*\text{-peer}}} = \arg\min_f \hat{R}_{\mathbb{1}_{\alpha^*\text{-peer}},\tilde{D}}(f)$ converges to f^*:

Theorem 5. *With probability at least* $1-\delta$,

$$R_{\mathcal{D}}(\hat{f}^*_{\mathbb{1}_{\alpha^*\text{-peer}}}) - R^* \leq \frac{1+\alpha^*}{1-e_{-1}-e_{+1}}\sqrt{\frac{2\log 2/\delta}{N}}.$$

4.5. Calibration and Generalization

So far our results focused on minimizing 0-1 losses, which is hard in practice. We provide evidence of ℓ_{peer}'s, and $\ell_{\alpha\text{-peer}}$'s in general, calibration and convexity for a generic and differentiable calibrated loss. We consider a ℓ that is classification calibrated, convex and L-Liptchitz.

Classification calibration describes the property that the excess risk when optimizing using a loss function ℓ would also guarantee a bound on the excessive 0-1 loss:

Definition 1. *ℓ is classification calibrated if there $\exists$ a convex, invertible, nondecreasing transformation Ψ_ℓ with $\Psi_\ell(0) = 0$ s.t. $\Psi_\ell(R_{\mathcal{D}}(\tilde{f}) - R^*) \leq R_{\ell,\mathcal{D}}(\tilde{f}) - \min_f R_{\ell,\mathcal{D}}(f), \forall \tilde{f}$.*

Denote $f^*_\ell \in \arg\min_f R_{\ell,\mathcal{D}}(f)$. Below we provide sufficient conditions for $\ell_{\alpha\text{-peer}}$ to be calibrated.

Theorem 6. *$\ell_{\alpha\text{-peer}}$ is classification calibrated when either of the following two conditions holds: (1) $\alpha = 1$ (i.e., $\ell_{\alpha\text{-peer}} = \ell_{peer}$), $p = 0.5$, and f^*_ℓ satisfies the following: $\mathbb{E}[\ell(f^*_\ell(X), -Y)] \geq \mathbb{E}[\ell(f(X), -Y)]$, $\forall f$. (2) $\alpha < 1, \max\{e_{+1}, e_{-1}\} < 0.5$, $\ell''(t,y) = \ell''(t,-y), \forall t$, and $\alpha(1-2p)(1-e_{+1}-e_{-1}) = (1-\alpha)(e_{+1}-e_{-1})$.*

(1) states that f^*_ℓ not only achieves the smallest risk over the clean distribution (X,Y) but also performs the worst on the "opposite" distribution with flipped labels $-Y$. (2) $\ell''(t,y) = \ell''(t,-y)$ is satisfied by some common loss function, such as square and logistic losses, as noted in (Natarajan et al., 2013),

Under the calibration condition, and denote the corresponding calibration transformation function for $\ell_{\alpha\text{-peer}}$ as $\Psi_{\ell_{\alpha\text{-peer}}}$.

Denote by

$$\hat{f}^*_{\ell_{\alpha\text{-peer}}} = \arg\min_{f\in\mathcal{F}} \hat{R}_{\ell_{\alpha\text{-peer}},\tilde{D}}(f) := \frac{1}{N}\sum_{n=1}^{N} \ell_{\alpha\text{-peer}}(f(x_n), \tilde{y}_n).$$

Consider a bounded ℓ with $\bar{\ell}, \underline{\ell}$ denoting its max and min value. We have the following generalization bound:

Theorem 7. *With probability at least* $1-\delta$:

$$R_{\mathcal{D}}(\hat{f}^*_{\ell_{\alpha^*\text{-peer}}}) - R^* \leq \frac{1}{1-e_{-1}-e_{+1}} \cdot \Psi^{-1}_{\ell_{\alpha^*\text{-peer}}}\Big(\min_{f\in\mathcal{F}} R_{\ell_{\alpha^*\text{-peer}},\tilde{\mathcal{D}}}(f) - \min_f R_{\ell_{\alpha^*\text{-peer}},\tilde{\mathcal{D}}}(f) + 4(1+\alpha^*)L\cdot\Re(\mathcal{F}) + 2\sqrt{\frac{\log 4/\delta}{2N}}\left(1+(1+\alpha^*)(\bar{\ell}-\underline{\ell})\right)\Big)$$

where $\Re(\mathcal{F})$ is Rademacher complexity of $\mathcal{F}$.

4.6. Convexity

In experiments, we use neural networks which are more robust to non-convex loss functions. Nonetheless, despite the fact that $\ell_{\alpha\text{-peer}}(\cdot)$ is not convex in general, Lemma 5 in (Natarajan et al., 2013) informs us that as long as $\hat{R}_{\ell_{\alpha\text{-peer}},\tilde{D}}(f)$ is close to some convex function, mirror gradient type of algorithms will converge to a small neighborhood of the optimal point when performing ERM with $\ell_{\alpha\text{-peer}}$. A natural candidate for this convex function is the expectation of $\hat{R}_{\ell_{\alpha\text{-peer}},\tilde{D}}(f)$ as $\hat{R}_{\ell_{\alpha\text{-peer}},\tilde{D}}(f) \to R_{\ell_{\alpha\text{-peer}},\tilde{\mathcal{D}}}(f)$ when $N \to \infty$.

Lemma 4. *When $\alpha < 1, \max\{e_{+1}, e_{-1}\} < 0.5$, $\ell''(t,y) = \ell''(t,-y), \forall t$, and $\alpha(1-2p)(1-e_{+1}-e_{-1}) = (1-\alpha)(e_{+1}-e_{-1})$, $R_{\ell_{\alpha\text{-peer}},\tilde{\mathcal{D}}}(f)$ is convex.*

This is the same condition as specified in (2) of Thm. 6.

5. Experiments

We implemented a two-layer ReLU Multi-Layer Perceptron (MLP) for classification tasks on 10 UCI Benchmarks and applied our peer loss to update their parameters. We show the robustness of peer loss with increasing rates of label noise on 10 real-world datasets. We compare the performance of our peer loss based method with surrogate loss method (Natarajan et al., 2013) (unbiased loss correction with known error rates), symmetric loss method (Ghosh et al., 2015), DMI (Xu et al., 2019), C-SVM (Liu et al., 2003) and PAM (Khardon & Wachman, 2007), which are state-of-the-art methods for dealing with random binary-classification noise, as well as a neural network baseline solution with binary cross entropy loss (NN). We use a cross-validation set to tune the parameters specific to the algorithms. For surrogate loss, we use the true e_{-1} and e_{+1} instead of learning them separately. Thus, surrogate loss could be considered a favored and advantaged baseline

Task		With Prior Equalization $p = 0.5$					Without Prior Equalization $p \neq 0.5$				
(d, N_+, N_-)	e_{-1}, e_{+1}	Peer	Surr	Symm	DMI	NN	Peer	Surr	Symm	DMI	NN
Twonorm (20,3700,3700)	0.1, 0.3	**0.977**	**0.968**	**0.969**	**0.974**	**0.964**	**0.977**	**0.968**	**0.969**	**0.974**	**0.964**
	0.2, 0.4	**0.976**	0.919	**0.959**	**0.966**	0.911	**0.976**	0.919	**0.959**	**0.966**	0.911
	0.4, 0.4	**0.973**	0.934	**0.958**	0.936	0.883	**0.973**	0.934	**0.958**	0.936	0.883
Splice (60,1527,1648)	0.1, 0.3	**0.919**	0.878	0.851	0.875	0.811	**0.925**	0.885	0.868	0.889	0.809
	0.2, 0.4	**0.901**	0.832	0.757	0.801	0.714	**0.912**	0.84	0.782	0.81	0.725
	0.4, 0.4	**0.819**	0.754	0.657	0.66	0.626	**0.822**	0.755	0.674	0.647	0.601
Diabetes (8,268,500)	0.1, 0.3	**0.833**	0.78	0.777	0.797	0.756	**0.856**	0.802	0.803	0.83	0.75
	0.2, 0.4	**0.755**	0.681	0.634	0.682	0.596	**0.739**	0.705	0.695	0.707	0.672
	0.4, 0.4	**0.719**	0.645	0.619	0.637	0.551	0.651	**0.685**	**0.68**	0.633	0.583
German (23,300,700)	0.1, 0.3	**0.639**	0.563	0.507	0.529	0.519	**0.727**	0.645	**0.709**	0.666	0.648
	0.2, 0.4	**0.664**	0.59	0.6	0.618	0.572	**0.676**	**0.681**	0.537	0.573	0.535
	0.4, 0.4	**0.606**	0.55	0.573	0.573	0.556	**0.654**	0.632	0.549	0.611	0.553
Waveform (21,1647,3353)	0.1, 0.3	**0.89**	**0.895**	**0.892**	0.856	0.868	**0.893**	**0.898**	**0.883**	0.785	0.863
	0.2, 0.4	**0.881**	**0.89**	0.828	0.835	0.81	**0.884**	**0.884**	0.745	0.761	0.837
	0.4, 0.4	**0.87**	**0.866**	**0.867**	0.773	0.835	**0.853**	**0.852**	**0.852**	0.672	0.828
Image (18,1320,990)	0.1, 0.3	**0.906**	**0.9**	**0.89**	0.87	0.909	**0.943**	0.909	0.897	0.811	**0.93**
	0.2, 0.4	0.836	**0.862**	0.719	**0.845**	0.832	0.672	0.755	0.722	**0.86**	0.599
	0.4, 0.4	0.741	0.72	**0.788**	0.763	0.732	**0.806**	**0.803**	**0.823**	0.762	0.8

Table 1. Experiment results on 6 UCI Benchmarks (The full table of all details on 10 UCI Benchmarks are deferred to Appendix; N_+, N_- are the numbers of positive and negative samples). Surr: surrogate loss method (Natarajan et al., 2013); DMI: (Xu et al., 2019); Symm: symmetric loss method (Ghosh et al., 2015). Entries within 2% from the best in each row are highlighted in bold. All results are averaged across 8 random seeds. Neural-network-based methods (Peer, Surrogate, NN, Symmetric, DMI) use the same hyper-parameters.

method. Accuracy of a classification algorithm is defined as the fraction of examples in the test set classified correctly with respect to the clean and true label. For given e_{+1} and e_{-1}, labels of the training data are flipped accordingly.

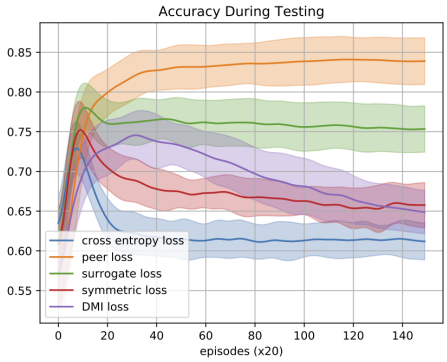

Figure 2. Accuracy on test set during training. Splice ($e_{-1} = 0.4$, $e_{+1} = 0.4$). More examples can be found in Appendix.

A subset of the experiment results is shown in Table 1. A full table with all details can be found in the Appendix. *Equalized Prior* means that we balance the dataset to guarantee $p = 0.5$. For this case we used ℓ_{peer} (i.e., $\alpha = 1$ as in $\ell_{\alpha\text{-peer}}$). For $p \neq 0.5$, we use validation dataset (still with noisy labels) to tune α. Our method is competitive across all datasets and is even able to outperform the surrogate loss method with access to the true noise rates in a number of datasets, as well as the symmetric loss functions (which does not require the knowledge of noise rates when error rates are symmetric) and the recently proposed information theoretical loss (Xu et al., 2019). Figure 2 shows that peer loss can prevent over-fitting when facing noisy labels.

A closer look at our decision boundary To have a better understanding of peer loss, we visualize the decision boundary returned by peer loss with a 2D synthetic experiment: the outer circle of randomly places points correspond to one class and the inner one is the other class. From Figure 3 we observe that when using cross entropy for training, the decision boundary is sharp on clean data but becomes much less so on noisy data (we have more examples with higher noise rate in the Appendix). Peer loss returns sharp boundaries even under a high noise rate (Figure 4).

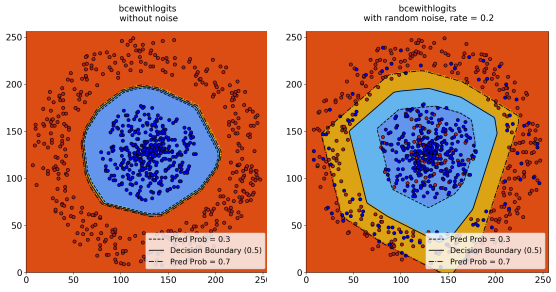

Figure 3. Decision boundary for cross entropy. Left: trained on clean data. Right: trained on noisy labels, $e_{+1} = e_{-1} = 0.2$.

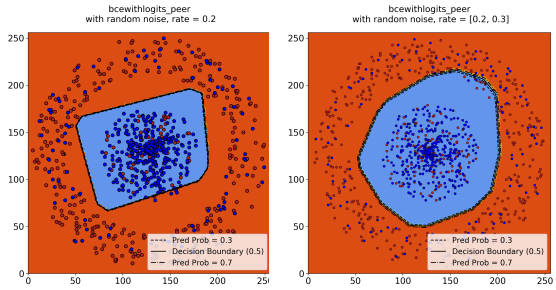

Figure 4. Decision boundary for peer loss. Left: $e_{+1} = e_{-1} = 0.2$. Right: $e_{+1} = 0.2, e_{-1} = 0.3$.

Preliminary results on multi-class classification We provide preliminary results on CIFAR-10 (Krizhevsky et al., 2009) in Table 2. We followed the setup in (Xu et al., 2019) and used ResNet (He et al., 2016) as the underlying optimization solution. However, different from settings in (Xu et al., 2019) where label noise only exists between specific class pairs, our noise is universal across classes. For each class, we flip the label to any other label with a probability of $\epsilon/9$, where ϵ is the error rate and 9 is the number of other classes. We do show peer loss is competitive against cross entropy and DMI (Xu et al., 2019).

Model	Error Rate $\epsilon = 0.2$	Error Rate $\epsilon = 0.4$
cross entropy	86.67	82.09
DMI (Xu et al., 2019)	85.11	81.67
Peer Loss	**87.72**	**83.81**

Table 2. Accuracy on CIFAR-10.

6. Conclusion and Discussion

This paper introduces peer loss, a family of loss functions that enables training a classifier over noisy labels, but without using explicit knowledge of the noise rates of labels.

Peer loss had made the assumption that label noise is homogeneous across training data instances. Future extensions of this work includes extension to instance based (Cheng et al., 2020; Xia et al., 2020) and margin based (Amid et al., 2019) label noise. We are also interested in exploring the application of peer loss in differentially private ERM (Chaudhuri et al., 2011), as well as in semi-supervised learning.

Proof for Lemma 2

Proof. We sketch the main steps. We denote by $X_{n_1}, \tilde{Y}_{n_2}$ the random variable corresponding to the peer samples $x_{n_1}, \tilde{y}_{n_2}$.

First we have

$$\mathbb{E}[\ell_{\text{peer}}(f(X), \tilde{Y})] = \mathbb{E}[\ell(f(X), \tilde{Y})] - \mathbb{E}[\ell(f(X_{n_1}), \tilde{Y}_{n_2})] \tag{7}$$

Consider the two terms on the RHS separately.

$$\begin{aligned}
&\mathbb{E}[\ell(f(X), \tilde{Y})] \\
=&\mathbb{E}_{X,Y=-1}\big[\mathbb{P}(\tilde{Y} = -1|Y = -1) \cdot \ell(f(X), -1) \\
&\quad + \mathbb{P}(\tilde{Y} = +1|Y = -1) \cdot \ell(f(X), +1)\big] \\
&+ \mathbb{E}_{X,Y=+1}\big[\mathbb{P}(\tilde{Y} = +1|Y = +1) \cdot \ell(f(X), +1) \\
&\quad + \mathbb{P}(\tilde{Y} = -1|Y = +1) \cdot \ell(f(X), -1)\big] \\
&\qquad\qquad \text{(Independence between } \tilde{Y} \text{ and } X \text{ given } Y) \\
=&\mathbb{E}_{X,Y=-1}\big[(1 - e_{-1})\ell(f(X), -1) + e_{-1}\ell(f(X), +1)\big] \\
+&\mathbb{E}_{X,Y=+1}\big[(1 - e_{+1})\ell(f(X), +1) + e_{+1}\ell(f(X), -1)\big]
\end{aligned} \tag{8}$$

The above is done mostly via law of total probability and using the assumption that $\tilde{Y}$ is conditionally (on Y) independent of X. Subtracting and adding $e_{+1} \cdot \ell(f(X), -1)$ and $e_{-1} \cdot \ell(f(X), +1)$ to the two expectation terms separately we have

$$\begin{aligned}
\text{Eqn. (8)} =&\mathbb{E}_{X,Y=-1}\big[(1 - e_{-1} - e_{+1}) \cdot \ell(f(X), -1) \\
&\quad + e_{+1} \cdot \ell(f(X), -1) + e_{-1} \cdot \ell(f(X), +1)\big] \\
&\quad + \mathbb{E}_{X,Y=+1}\big[(1 - e_{-1} - e_{+1}) \cdot \ell(f(X), +1) \\
&\quad + e_{-1} \cdot \ell(f(X), +1) + e_{+1} \cdot \ell(f(X), -1)\big] \\
=&(1 - e_{-1} - e_{+1}) \cdot \mathbb{E}_{X,Y}\big[\ell(f(X), Y)\big] \\
&\quad + \mathbb{E}_X\big[e_{+1} \cdot \ell(f(X), -1) + e_{-1} \cdot \ell(f(X), +1)\big]
\end{aligned}$$

And consider the second term:

$$\begin{aligned}
&\mathbb{E}[\ell(f(X_{n_1}), \tilde{Y}_{n_2})] \\
=&\mathbb{E}_X[\ell(f(X), -1)] \cdot \mathbb{P}(\tilde{Y} = -1) \\
&\quad + \mathbb{E}_X[\ell(f(X), +1)] \cdot \mathbb{P}(\tilde{Y} = +1) \\
&\qquad\qquad \text{(Independence between } n_1 \text{ and } n_2) \\
=&\mathbb{E}_X\big[(e_{+1} \cdot p + (1 - e_{-1})(1 - p)) \cdot \ell(f(X), -1) \\
&\quad + ((1 - e_{+1})p + e_{-1}(1 - p)) \cdot \ell(f(X), +1)\big] \\
&\qquad\qquad \text{(Expressing } \mathbb{P}(\tilde{Y}) \text{ using } p \text{ and } e_{+1}, e_{-1}) \\
=&\mathbb{E}_X\big[(1 - e_{-1} - e_{+1})(1 - p) \cdot \ell(f(X), -1) \\
&\quad + (1 - e_{-1} - e_{+1})p \cdot \ell(f(X), +1)\big] \\
&\quad + \mathbb{E}_X\big[(e_{+1} \cdot p + e_{+1}(1 - p)) \cdot \ell(f(X), -1) \\
&\quad + (e_{-1}(1 - p) + e_{-1}p) \cdot \ell(f(X), +1)\big] \\
=&(1 - e_{-1} - e_{+1}) \cdot \mathbb{E}[\ell(f(X_{n_1}), Y_{n_2})] \\
&\quad + \mathbb{E}_X\big[e_{+1} \cdot \ell(f(X), -1) + e_{-1} \cdot \ell(f(X), +1)\big]
\end{aligned}$$

Subtracting the first and second term on RHS of Eqn. (7):

$$\begin{aligned}
\mathbb{E}[\ell_{\text{peer}}(f(X), \tilde{Y})] &= \mathbb{E}[\ell(f(X), \tilde{Y})] - \mathbb{E}[\ell(f(X_{n_1}), \tilde{Y}_{n_2})] \\
&= (1 - e_{-1} - e_{+1}) \cdot \mathbb{E}[\ell_{\text{peer}}(f(X), Y)]
\end{aligned} \tag{9}$$

□

Acknowledgement

Yang Liu would like to thank Yiling Chen for inspiring early discussions on this problem. The authors thank Tongliang Liu, Ehsan Amid and Manfred Warmuth for constructive comments and conversations, and Nontawat Charoenphakdee for his comments on related works. The authors also would like to thank Xingyu Li, Zhaowei Zhu and Jiaheng Wei for detailed discussions, suggestions and help with generating Figures 3 and 4.

This work is partially funded by the Defense Advanced Research Projects Agency (DARPA) and Space and Naval Warfare Systems Center Pacific (SSC Pacific) under Contract No. N66001-19-C-4014. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of DARPA, SSC Pacific or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

Amid, E., Warmuth, M. K., Anil, R., and Koren, T. Robust bi-tempered logistic loss based on bregman divergences. In *Advances in Neural Information Processing Systems*, pp. 15013–15022, 2019.

Ben-David, S., Pál, D., and Shalev-Shwartz, S. Agnostic online learning. In *COLT 2009*.

Bylander, T. Learning linear threshold functions in the presence of classification noise. In *Proceedings of the seventh annual conference on Computational learning theory*, pp. 340–347. ACM, 1994.

Cesa-Bianchi, N., Dichterman, E., Fischer, P., Shamir, E., and Simon, H. U. Sample-efficient strategies for learning in the presence of noise. *Journal of the ACM*, 1999.

Cesa-Bianchi, N., Shalev-Shwartz, S., and Shamir, O. Online learning of noisy data. *IEEE Transactions on Information Theory*, 57(12):7907–7931, 2011.

Charoenphakdee, N., Lee, J., and Sugiyama, M. On symmetric losses for learning from corrupted labels. In *International Conference on Machine Learning*, 2019.

Chaudhuri, K., Monteleoni, C., and Sarwate, A. D. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(3), 2011.

Cheng, J., Liu, T., Ramamohanarao, K., and Tao, D. Learning with bounded instance-and label-dependent label noise. *ICML, arXiv:1709.03768*, 2020.

Dasgupta, A. and Ghosh, A. Crowdsourced judgement elicitation with endogenous proficiency. In *Proceedings of the 22nd international conference on World Wide Web*, pp. 319–330. International World Wide Web Conferences Steering Committee, 2013.

Du Plessis, M. C., Niu, G., and Sugiyama, M. Clustering unclustered data: Unsupervised binary labeling of two datasets having different class balances. In *2013 Conference on Technologies and Applications of Artificial Intelligence*, pp. 1–6. IEEE, 2013.

Ghosh, A., Manwani, N., and Sastry, P. Making risk minimization tolerant to label noise. *Neurocomputing*, 2015.

Ghosh, A., Kumar, H., and Sastry, P. Robust loss functions under label noise for deep neural networks. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

Goldberger, J. and Ben-Reuven, E. Training deep neural-networks using a noise adaptation layer. 2016.

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 *Advances in neural information processing systems*, 2018.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Jenni, S. and Favaro, P. Deep bilevel learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 618–633, 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. *arXiv preprint arXiv:1712.05055*, 2017.

Khardon, R. and Wachman, G. Noise tolerant variants of the perceptron algorithm. *J. Mach. Learn. Res.*, 8:227–248, May 2007.

Kong, Y. and Schoenebeck, G. Water from two rocks: Maximizing the mutual information. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pp. 177–194. ACM, 2018.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. Technical report, 2009.

Liu, B., Dai, Y., Li, X., Lee, W. S., and Yu, P. S. Building text classifiers using positive and unlabeled examples. In *Proceedings of the Third IEEE International Conference on Data Mining*, ICDM '03, pp. 179–, Washington, DC, USA, 2003. IEEE Computer Society.

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

Liu, Y. and Chen, Y. Machine Learning aided Peer Prediction. *ACM EC*, June 2017.

Liu, Y. and Guo, H. Peer loss functions: Learning from noisy labels without knowing noise rates. In *International Conference on Machine Learning*, 2020.

Lu, N., Niu, G., Menon, A. K., and Sugiyama, M. On the minimal supervision for training any binary classifier from only unlabeled data. *arXiv preprint arXiv:1808.10585*, 2018.

Manwani, N. and Sastry, P. Noise tolerance under risk minimization. *IEEE transactions on cybernetics*, 43(3): 1146–1151, 2013.

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

Miller, N., Resnick, P., and Zeckhauser, R. Eliciting informative feedback: The peer-prediction method. *Management Science*, 51(9):1359 –1373, 2005.

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

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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1944–1952, 2017.

Prelec, D. A bayesian truth serum for subjective data. *Science*, 306(5695):462–466, 2004.

Radanovic, G. and Faltings, B. A robust bayesian truth serum for non-binary signals. In *Proceedings of the 27th AAAI Conference on Artificial Intelligence*, 2013.

Scott, C. A rate of convergence for mixture proportion estimation, with application to learning from noisy labels. In *AISTATS*, 2015.

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

Shnayder, V., Agarwal, A., Frongillo, R., and Parkes, D. C. Informed truthfulness in multi-task peer prediction. In *Proceedings of the 2016 ACM Conference on Economics and Computation*, pp. 179–196. ACM, 2016.

Song, H., Kim, M., and Lee, J.-G. Selfie: Refurbishing unclean samples for robust deep learning. In *International Conference on Machine Learning*, pp. 5907–5915, 2019.

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

Sukhbaatar, S. and Fergus, R. Learning from noisy labels with deep neural networks. *arXiv preprint arXiv:1406.2080*, 2(3):4, 2014.

Van Rooyen, B., Menon, A., and Williamson, R. C. Learning with symmetric label noise: The importance of being unhinged. In *Advances in Neural Information Processing Systems*, pp. 10–18, 2015a.

Van Rooyen, B., Menon, A. K., and Williamson, R. C. An average classification algorithm. *arXiv preprint arXiv:1506.01520*, 2015b.

Witkowski, J. and Parkes, D. A robust bayesian truth serum for small populations. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence*, AAAI '12, 2012.

Witkowski, J., Bachrach, Y., Key, P., and Parkes, D. C. Dwelling on the Negative: Incentivizing Effort in Peer Prediction. In *Proceedings of the 1st AAAI Conference on Human Computation and Crowdsourcing (HCOMP'13)*, 2013.

Xia, X., Liu, T., Han, B., Wang, N., Gong, M., Liu, H., Niu, G., Tao, D., and Sugiyama, M. Parts-dependent label noise: Towards instance-dependent label noise, 2020.

Xiao, T., Xia, T., Yang, Y., Huang, C., and Wang, X. Learning from massive noisy labeled data for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

Xu, Y., Cao, P., Kong, Y., and Wang, Y. L_dmi: An information-theoretic noise-robust loss function. *NeurIPS, arXiv:1909.03388*, 2019.

Yi, K. and Wu, J. Probabilistic end-to-end noise correction for learning with noisy labels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7017–7025, 2019.

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

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