Learning from Noisy Similar and Dissimilar Data

Soham Dan ✉[1] ⋆, Han Bao[2,3], and Masashi Sugiyama[2,3]

[1] University of Pennsylvania
[2] RIKEN Center for Advanced Intelligence Project [3] The University of Tokyo
sohamdan@seas.upenn.edu, tsutsumi@ms.k.u-tokyo.ac.jp,
sugi@k.u-tokyo.ac.jp

Abstract. With the widespread use of machine learning for classification, it becomes increasingly important to be able to use weaker kinds of supervision for tasks in which it is hard to obtain standard labeled data. One such kind of supervision is provided *pairwise* in the form of Similar (S) pairs (if two examples belong to the same class) and Dissimilar (D) pairs (if two examples belong to different classes). This kind of supervision is realistic in privacy-sensitive domains. Although the basic version of this problem has been studied recently, it is still unclear how to learn from such supervision under *label noise*, which is very common when the supervision is, for instance, crowd-sourced. In this paper, we close this gap and demonstrate how to learn a classifier from noisy S and D labeled pairs. We perform a detailed investigation of this problem under two realistic noise models and propose two algorithms to learn from noisy SD data. We also show important connections between learning from such pairwise supervision data and learning from ordinary class-labeled data. Finally, we perform experiments on synthetic and real-world datasets and show our noise-informed algorithms outperform existing baselines in learning from noisy pairwise data.

Keywords: Classification · Pairwise Supervision · Noisy Supervision

1 Introduction

In the standard supervised learning framework, a classifier is trained with labeled data points, which are usually collected through human annotation. While collecting labeled data points is the traditional way to apply supervised classification, *pairwise comparison* is often more appealing for human decision making [10], where annotators are requested to compare two instances and give relative relationships between them; e.g., which instance has stronger stimulus, whether two instances belong to the same category, and so on. This is partly because (a) decision makers tend to be subjective at directly choosing a single hypothesis, [3] and (b) decision makers are often biased about picking an opinion.[4]

[⋆] Work done during an internship at RIKEN-AIP

[3] [24] has studied a relationship between relative comparison and a single hypothesis on stimuli, which is known as the law of comparative judgement.

[4] This bias is known as social desirability bias [9]; questionees are unconsciously led to a socially desirable opinion when they are asked to reveal their opinions in a direct

This relative ease of making pairwise comparisons over direct point-wise labeling has inspired several successful large-scale annotation frameworks, for example, crowd-clustering [11, 26]. There are broadly two ways to incorporate pairwise comparisons for identifying the latent classes of data:
(1) Semi-Supervised Clustering based methods [5]:, which utilizes pairwise supervision indicating whether two instances belong to the same cluster or not (known as must-link and cannot-link constraints), guiding clustering as decision makers desire. This class of methods suffer from dataset-dependent assumptions.
(2) Empirical Risk Minimization (ERM) based methods: [1, 23] which trains an inductive classifier from the pairwise comparisons thereby establishing a connection to standard supervised learning in the ERM framework. These methods outperform semi-supervised clustering based methods because it does not make similar assumptions as the latter. In this paper, our primary focus is on the second class of methods, aiming to learn inductive classifiers from pairwise data, and we shall see in Sec. 5 they outperform the first class of methods empirically. It is important to note that both methods assume that the pairwise comparisons are noise-free, i.e.,instances marked similar are indeed from the same class.
While learning from pairwise comparisons has been highly successful [13, 10, 14, 8, 21], it is sensitive to the quality of the annotations. Large-scale frameworks like crowd-clustering are especially prone to noisy annotations [26]. Existing techniques to learn an inductive classifier from noisy labels [19, 20, 15, 12] are inapplicable in this setting, since they only work on pointwise, class-labeled data. Moreover, there lacks a systematic characterization of the kinds of noise that might arise in pairwise-annotated data. We aim to bridge this gap by characterizing the two unique types of errors that arise in this setting, as depicted in Figure 1. The first error results from *pairing corruption*: some pairs of instances are hard to identify whether they belong to the same category or not. The second error is from *labeling corruption*: labels of some instances are intrinsically ambiguous and thus, subsequent pairwise comparison is also affected. Each of these situations give rise to a specific noise model for pairwise supervision.

In this paper, we thoroughly investigate classification with noisy pairwise supervision, where the noise is present in pairwise comparison and follows either pairing corruption or labeling corruption, and provide two distinct strategies to deal with this problem. In the first strategy, we introduce a corrected loss function, which induces an unbiased estimator of the classification risk in the presence of noise for pairwise data. Subsequently, a classifier can be obtained through the minimization of the corrected loss. This extends previous approaches [19, 20] to the pairwise setting. The second strategy is motivated from the insight that the Bayes classifier of the classification risk under the noise-free distribution corresponds to that of the weighted risk under the noisy distribution. This extends cost-sensitive classification [7, 22] to the pairwise setting. Each of these strategies can handle both the pairwise noise models.
We make the following contributions in this paper:

way. Such a tendency is observed especially in answering their sensitive matters such as criminal records.

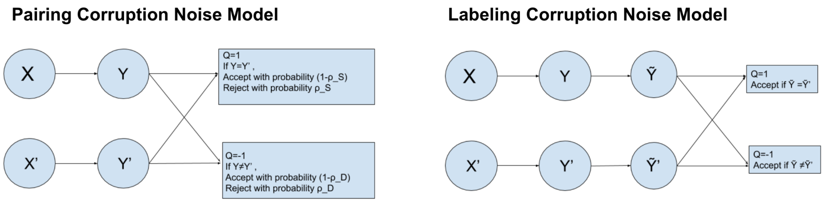
Pairing Corruption Noise Model
X
Y
X'
Y'
Q=1
If Y=Y' ,
Accept with probability (1-ρ_S)
Reject with probability ρ_S
Q=-1
If Y≠Y' ,
Accept with probability (1-ρ_D)
Reject with probability ρ_D
Labeling Corruption Noise Model
X
Y
Ỹ
X'
Y'
Ỹ'
Q=1
Accept if Ỹ =Ỹ'
Q=-1
Accept if Ỹ ≠Ỹ'

Fig. 1: The two noise models. Q indicates whether the pair is similar ($Q = 1$) or dissimilar ($Q = -1$)

- We provide two distinct, realistic data generating scenarios for SD data: the pairing corruption noise model and the labeling corruption noise model.
- We provide two algorithms based on loss correction and weighted classification which can be applied to either data generating scenario mentioned.
- We theoretically analyze performance bounds of our algorithms and provide two new performance bounds for the noise-free SD learning problem [23].
- We perform extensive experiments on various datasets to show that the proposed algorithms work well in practice and outperform existing methods.

2 Problem Setup

Let $\mathcal{X}$ denote the instance space, $\mathcal{Y} = \{+1, -1\}$ the label space and $\mathcal{Z}$ the underlying distribution over $(\mathcal{X}, \mathcal{Y})$. We want to perform well with respect to $\mathcal{Z}$, i.e., the test data is drawn from $\mathcal{Z}$ and we want to minimize the risk of f (a real valued decision function) w.r.t. the 0-1 loss:

$$R_{\mathcal{Z}}(f) = \mathbb{E}_{(x,y)\sim\mathcal{Z}}[\mathbb{1}_{\{\text{sign}(f(x))\neq y\}}], \tag{1}$$

where $\mathbb{1}_{\{.\}}$ denotes the indicator function, $f \in \mathcal{F}$ and $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$ is a hypothesis class. However, we assume that due to the domain constraints we are unable to procure direct class-labeled data from $\mathcal{Z}$ and only have access to pairwise supervision—whether a pair of instances (X, X') is from the same class ($Y = Y'$) or from different classes ($Y \neq Y'$). There is a latent variable Q dictating whether the pair is similar ($Q = 1$) or dissimilar ($Q = -1$). Our training dataset, $\mathcal{D} = \{X_i, X'_i, Q_i\}_{i=1}^{n}$, consists of n noisy similar or dissimilar pairs which are generated from one of the following noise models that reflect what we may expect in real-world data. Both of them are illustrated in Figure 1.

Noise Model 1: Pairing Corruption: This model is motivated by the following scenario—imagine crowd-workers are given a pool of instances drawn from $\mathcal{X}$ and they annotate pairs (X, X') as $Q = 1$ if they believe $Y = Y'$ and $Q = -1$ otherwise. Since they are not experts, they often make mistakes in this process and sometimes assigns $Q = -1$ when it should have been $Q = 1$ and vice versa. Formally, this model comprises of the following steps:

1. Two samples (X, Y),(X', Y') are drawn from the underlying distribution $\mathcal{Z}$.

2. If $Y = Y'$, this pair of samples is labeled as similar ($Q = 1$) with probability $1 - \rho_{\mathrm{S}}$ and if $Y \neq Y'$ this pair is labeled as dissimilar ($Q = -1$) with probability $1 - \rho_{\mathrm{D}}$, where ρ_{S} and ρ_{D} are the noise rate for similar (S) and dissimilar (D) data respectively and $0 \leq \rho_{\mathrm{S}}, \rho_{\mathrm{D}} \leq 1$.

In this noise model, the S and D samples are drawn from mixtures of the true S and D distributions: $P(Q = 1|Y = Y') = 1-\rho_{\mathrm{S}}$ and $P(Q = -1|Y \neq Y') = 1-\rho_{\mathrm{D}}$. We assume that $\rho_{\mathrm{S}} + \rho_{\mathrm{D}} < 1$, without which it is impossible to learn a classifier.
Noise Model 2: Labeling Corruption: Consider that we are dealing with a privacy sensitive domain where responders do not want to reveal their individual labels. In such cases, responders may intentionally reveal wrong labels. Hence, the pointwise labels that we obtain are intrinsically noisy. A moderator converts the pointwise data (X, Y) to pairwise data $(X, X', Q = \pm 1)$, to preserve privacy. Formally, there are the following steps in this noise model:

1. Two samples (X, Y),(X', Y') are drawn from the underlying distribution $\mathscr{Z}$.
2. Then the labels are flipped with probability $\rho_{\pm}$ (this is class conditioned: if a sample originally has label $+1$ it is flipped to label -1 with probability ρ_+ and respectively, ρ_- for the other case). Formally, this can be expressed as $P(\tilde{Y} = +1|Y = -1) = \rho_-$ and $P(\tilde{Y} = -1|Y = +1) = \rho_+$.
3. Then in the next step the similar or dissimilar labels are assigned (there is assumed to be no noise in this step since the moderator is an expert).

Thus, $P(Q = 1|\tilde{Y} = \tilde{Y'}) = 1$ and $P(Q = -1|\tilde{Y} \neq \tilde{Y'}) = 1$. Again, in this noise model we assume $\rho_+ + \rho_- < 1$. In the following, we derive the conditional density functions separately for the noise-free, pairing noise and labeling noise scenarios and show how one can view the pairwise samples as pointwise samples.
Conditional density functions in the noise-free case: We consider the one-sample case of the problem of learning from noisy SD labels, which means only one training dataset of SD samples are drawn (as opposed to the two sample case of separate S and D data) from a joint distribution $P(x, x', Q)$. Let us consider the simpler, noise-free scenario first. We can write the joint distribution as:

$$P(x, x', Q) = P(x, x'|Q = 1)P(Q = 1) + P(x, x'|Q = -1)P(Q = -1).$$

If there were no noise corruption, S data would comprise of two positive instances or two negative instances and D data would comprise of one positive instance and one negative instance. Accordingly, we obtain the following conditional distributions for the S and D pairs as follows:

$$\begin{aligned}
P(x, x'|Q = 1) &= P(x, x'|y = y' = 1 \vee y = y' = -1), \\
&= \frac{\pi^2 P_+(x)P_+(x') + (1-\pi)^2 P_-(x)P_-(x')}{\pi^2 + (1-\pi)^2}, \\
P(x, x'|Q = -1) &= P(x, x'|y = 1, y' = -1 \vee y = -1, y' = 1), \\
&= \frac{\pi(1-\pi)P_+(x)P_-(x') + \pi(1-\pi)P_-(x)P_+(x')}{2\pi(1-\pi)}.
\end{aligned} \tag{2}$$

where $\pi = P(Y = +1)$ denotes the class prior, $P_+(X) = P(X|Y = +1)$ and $P_-(X) = P(X|Y = -1)$, $P(Q = 1) = \pi^2 + (1-\pi)^2$ and $P(Q = -1) = 2\pi(1-\pi)$.

One can further marginalize out x' and get the densities in terms of a single data point x. This view is important for our subsequent analysis and has been previously used in [1]. The implication is that we can now treat the SD data as pointwise data from the Similar (S) and Dissimilar (D) classes.

$$\begin{aligned} P_{\mathrm{S}}(x) &= P(x|Q=1) = \frac{\pi^2 P_+(x) + (1-\pi)^2 P_-(x)}{\pi^2 + (1-\pi)^2}, \\ P_{\mathrm{D}}(x) &= P(x|Q=-1) = \frac{P_+(x) + P_-(x)}{2}. \end{aligned} \tag{3}$$

Conditional density functions under pairwise noise: The above expressions are derived under the noise-free assumption. Now, we present the expression for the densities, $\tilde{P}_{\mathrm{S}}(x)$ and $\tilde{P}_{\mathrm{D}}(x)$, under each of the noise models presented above. The derivations follow from the graphical model in Fig. 1 and can be found in the Appendix 1.
For the pairing corruption model we get:

$$\tilde{P}_{\mathrm{S}}(x) = (1-\rho_{\mathrm{S}})P_{\mathrm{S}}(x) + \rho_{\mathrm{D}} P_{\mathrm{D}}(x), \quad \tilde{P}_{\mathrm{D}}(x) = \rho_{\mathrm{S}} P_{\mathrm{S}}(x) + (1-\rho_{\mathrm{D}})P_{\mathrm{D}}(x), \tag{4}$$

and for the labeling corruption model we get,

$$\begin{aligned} \tilde{P}_{\mathrm{S}}(x) &= \frac{(\pi(1-\rho_+)P_+(x) + (1-\pi)\rho_- P_-(x))\tilde{\pi}}{\tilde{\pi}^2 + (1-\tilde{\pi})^2} \\ &\quad + \frac{((1-\pi)(1-\rho_-)P_-(x) + \pi\rho_+ P_+(x))(1-\tilde{\pi})}{\tilde{\pi}^2 + (1-\tilde{\pi})^2}, \\ \tilde{P}_{\mathrm{D}}(x) &= \frac{(\pi\rho_+ P_+(x) + (1-\pi)\rho_- P_-(x))\tilde{\pi}}{2} \\ &\quad + \frac{(\pi(1-\rho_+)P_-(x) + (1-\pi)(1-\rho_-)P_+(x))(1-\tilde{\pi})}{2}, \\ &\text{where} \quad \tilde{\pi} = \pi(1-\rho_+) + (1-\pi)\rho_-. \end{aligned} \tag{5}$$

Let $\mathcal{Z}_Q$ denote the distribution over $(\mathcal{X}, Q)$ where $P(X|Q=1) = \tilde{P}_{\mathrm{S}}(X)$ and $P(X|Q=-1) = \tilde{P}_{\mathrm{D}}(X)$. Given the marginalized representations in (4) and (5), we can now think of the training data of n pairwise instances (drawn from either of the noise models) to be equivalent to a training data of $2n$ pointwise instances drawn from $\mathcal{Z}_Q$, i.e., $\mathcal{D} \triangleq \{X_i, X_i', Q_i\}_{i=1}^n \equiv \mathcal{D}' \triangleq \{X_i, Q_i\}_{i=1}^{2n} \sim \mathcal{Z}_Q$.

3 Loss Correction Approach

In this section, we present the first of our two proposed algorithms for learning from noisy pairwise data. In this method, we derive an unbiased estimator of the classification risk (Eq. 1) on noisy SD data, by modifying the standard loss function for binary classification. We assume that the noise rates (ρ_{S} and ρ_{D} for the pairing corruption model and $\rho_\pm$ for the labeling corruption model) are available beforehand, which are then used to obtain an unbiased estimator of the classification risk. [20] studies a technique for loss correction in the standard classification setting. We adapt this backward-correction technique to handle noisy SD data. A key step necessary to correct the loss is to write the posterior over Q in terms of the posterior over Y.

We now present the posterior SD probabilities in terms of the ordinary class posterior probabilities. The detailed derivation of the following equations can be found in Appendix 2. For the pairing corruption noise model we obtain:

$$\begin{aligned} P(Q=1|X) &= P(Y=1|X)[(1-\rho_{\mathrm{S}})\pi + \rho_{\mathrm{D}}(1-\pi)] \\ &\quad + P(Y=-1|X)[\rho_{\mathrm{D}}\pi + (1-\rho_{\mathrm{S}})(1-\pi)], \\ P(Q=-1|X) &= P(Y=-1|X)[(1-\rho_{\mathrm{D}})\pi + \rho_{\mathrm{S}}(1-\pi)] \\ &\quad + P(Y=1|X)[\rho_{\mathrm{S}}\pi + (1-\rho_{\mathrm{D}})(1-\pi)]. \end{aligned} \tag{6}$$

On the other hand, for the labeling corruption noise model we obtain:

$$\begin{aligned} P(Q=1|X) &= P(Y=1|X)[(1-\rho_{+})\tilde{\pi} + \rho_{-}(1-\tilde{\pi})] \\ &\quad + P(Y=-1|X)[\rho_{-}\tilde{\pi} + (1-\rho_{+})(1-\tilde{\pi})], \\ P(Q=-1|X) &= P(Y=-1|X)[(1-\rho_{-})\tilde{\pi} + \rho_{+}(1-\tilde{\pi})] \\ &\quad + P(Y=1|X)[\rho_{+}\tilde{\pi} + (1-\rho_{-})(1-\tilde{\pi})]. \end{aligned} \tag{7}$$

Thus, in both noise models we can express:
$P(Q=1|x) = \alpha_1 P(y=1|x) + \alpha_2 P(y=-1|x)$ and
$P(Q=-1|x) = \beta_1 P(y=1|x) + \beta_2 P(y=-1|x)$ for some coefficients $\alpha_1, \alpha_2, \beta_1, \beta_2$. It is noteworthy that the structure of the posterior probabilities are remarkably similar between the two noise models once we introduce the modified class prior $\tilde{\pi}$, defined in (5). Hence, the labeling corruption noise model can be interpreted as the following two-step data generating process: pointwise labels following the modified class prior $\tilde{\pi}$ are observed first, then pairwise labels are observed via the pairing corruption noise model with noise rates ρ_+ and ρ_-.
Having expressed the posterior probabilities of the noisy SD data in terms of the original class posteriors, we can adopt the technique of backward correction to construct the modified loss function such that, the minimizer of the expected risk with this new loss function over the noisy SD data is the same as the minimizer of the expected risk with the original loss function over $\mathcal{Z}$ (test distribution).
Let $\ell : \mathbb{R} \times \mathcal{Y} \rightarrow \mathbb{R}_{\geq 0}$ be a loss function such that $\ell(t, y)$ measures discrepancy between the prediction t and the target label y and let $T = \begin{bmatrix} \alpha_1 & \alpha_2 \\ \beta_1 & \beta_2 \end{bmatrix}$. If $\tilde{\ell}(t) = T^{-1}\ell(t)$ denotes the backward corrected loss, then $\ell(t) = \mathbb{E}[T\tilde{\ell}(t)]$. Note that the expectation is taken w.r.t Q. $\tilde{\ell}(t, Q)$ can be obtained as the first or the second row of $T^{-1}\ell(t)$ corresponding to $Q = +1$ or $Q = -1$ respectively.

Remark 1. Here we have assumed T is invertible which almost always holds in practice. However, if the condition number is large, i.e., T is almost singular, we can mix T (with an appropriate value of λ) with the matrix T' which corresponds to the noise-free SD case (obtained by setting noise rates to 0 in (6),(7); refer to

(13) below) before inverting it. This is essentially including a noise-free prior.

$$T \leftarrow T + \lambda T', \quad \text{where} \quad T' = \begin{bmatrix} \pi & 1-\pi \\ 1-\pi & \pi \end{bmatrix}$$

The following is the empirical $\tilde{\ell}$-risk on the observed pairwise training data $\mathcal{D}$ of n instances.

$$\hat{R}_{\tilde{\ell}}(f) = \frac{1}{n} \sum_{i=1}^{n} \tilde{\ell}(g(X_i, X_i'), Q_i), \tag{8}$$

where, $g \in \mathcal{G}$ is a real-valued decision function and $\mathcal{G} \subseteq \mathbb{R}^{\mathcal{X}} \times \mathbb{R}^{\mathcal{X}}$. Further, for ease of analysis we always deal with the pointwise view of the training dataset $\mathcal{D}'$ with $2n$ instances (refer to Sec. 2 above for more details). In this view, the empirical $\tilde{\ell}$-risk on the observed pointwise training data $\mathcal{D}'$ of $2n$ instances is:

$$\hat{R}_{\tilde{\ell}}(f) = \frac{1}{2n} \sum_{i=1}^{2n} \tilde{\ell}(f(X_i), Q_i), \tag{9}$$

where $f \in \mathcal{F}$ is a real-valued decision function ,$\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$ is a hypothesis class. We can use the corrected loss to train our classifier f on the noisy SD data directly by empirical risk minimization of (9).

Performance Bounds: Now we discuss the performance bounds for this approach. The following are some important notations used in the following results:

- $\hat{f} = \arg\min_{f \in \mathcal{F}} \hat{R}_{\tilde{\ell}}(f)$,
- $R_{\tilde{\ell}, \mathcal{Z}_Q}(f) = E_{(X,Q) \sim \mathcal{Z}_Q}[\tilde{\ell}(f(X), Q)]$,
- $R_{\ell, \mathcal{Z}}(f) = E_{(X,Y) \sim \mathcal{Z}}[\ell(f(X), Y)]$.

The empirical estimate of the risk is unbiased to $R_{\ell,\mathcal{Z}}(f)$ because $\tilde{\ell}$ is corrected appropriately. By performing ERM on the noisy SD data using the corrected loss, the empirical risk converges to the true risk on the standard class-labeled Positive(P)-Negative(N) data drawn from the underlying distribution $\mathcal{Z}$. Let L_Q be the Lipschitz constant of the loss $\tilde{\ell}$ in its first argument. Note that $L_Q \le \max\{\alpha_1, \alpha_2, \beta_1, \beta_2\}L$. Let $\mathcal{R}(\mathcal{F}, 2n)$ be the Rademacher complexity [4] of the function class $\mathcal{F}$ for the $2n$ noisy SD instances.

Further, for the following theorem we also need the notion of classification calibrated surrogate losses. If a surrogate loss $\ell(\cdot,\cdot)$ is calibrated, then convergence of the surrogate excess risk $R_{\ell,\mathcal{Z}}(f) - \min_f R_{\ell,\mathcal{Z}}(f)$ to zero implies convergence of the target excess risk $R_{\mathcal{Z}}(f) - \min_f R_{\mathcal{Z}}(f)$ to zero. For more details refer to [3].

Lemma 1. *For any $\delta > 0$, we have the following: with probability at least $1-\delta$,*

$$\max_{f \in \mathcal{F}} |\hat{R}_{\tilde{\ell}}(f) - R_{\tilde{\ell}, \mathcal{Z}_Q}(f)| \le 2L_Q \mathcal{R}(\mathcal{F}, 2n) + \sqrt{\frac{\log(1/\delta)}{4n}}. \tag{10}$$

The proof of Lemma 1 can be found in Appendix 3. We see that the generalization error of any f w.r.t. $\mathcal{Z}_Q$ vanishes asymptotically if $\mathcal{R}(\mathcal{F}, 2n)$ is moderately controlled as is the case for linear-in-parameter models [18].

Theorem 1. *For any $\delta > 0$, we have the following: with probability at least $1-\delta$,*

$$R_{\ell,\mathcal{Z}}(\hat{f}) \leq \min_{f\in\mathcal{F}} R_{\ell,\mathcal{Z}}(f) + 4L_Q\mathcal{R}(\mathcal{F},2n) + 2\sqrt{\frac{\log(1/\delta)}{4n}}. \tag{11}$$

If ℓ is classification calibrated [3], there exists a non-decreasing function ξ_ℓ with $\xi_\ell(0) = 0$ such that,

$$R_{\mathcal{Z}}(\hat{f}) - R^* \leq \xi_\ell^{-1}\left(\min_{f\in\mathcal{F}} R_{l,\mathcal{Z}}(f) - \min_{f} R_{l,\mathcal{Z}}(f) + 4L_Q\mathcal{R}(\mathcal{F},2n) + 2\sqrt{\frac{\log(1/\delta)}{4n}}\right). \tag{12}$$

Detailed proof is available in Appendix 4.
We see that the estimation error of $\hat{f}$ vanishes asymptotically if $\mathcal{R}(\mathcal{F},2n)$ is moderately controlled as is the case for linear-in-parameter models [18].
Special Case of Noise-Free SD Learning: When there is no noise $\rho_{\mathrm{S}} = \rho_{\mathrm{D}} = 0$ or $\rho_+ = \rho_- = 0$ and $\pi \neq 0.5$, for both the noise models,

$$T = \begin{bmatrix} \pi & 1-\pi \\ 1-\pi & \pi \end{bmatrix} \implies T^{-1} = \begin{bmatrix} \frac{\pi}{(2\pi-1)} & \frac{-(1-\pi)}{(2\pi-1)} \\ \frac{-(1-\pi)}{(2\pi-1)} & \frac{\pi}{(2\pi-1)} \end{bmatrix}. \tag{13}$$

We see this matches the loss function derived for noise-free SD learning in [23] (on setting X' as X and replacing $\pi_S E_{X_S}[1] = \pi_D E_{X_D}[1] = \frac{1}{2n}$):

$$\hat{R}_{\tilde{\ell}}(f) = \frac{1}{2n}\sum_{i=1}^{2n}[\mathcal{L}(f(X_i),Q_i)], \quad \text{where} \quad \mathcal{L}(z,t) = \frac{\pi}{2\pi-1}\ell(z,t) - \frac{1-\pi}{2\pi-1}\ell(z,-t). \tag{14}$$

Our analysis through the lens of loss correction provides an estimation error bound for noise-free SD learning as a special case of noisy SD learning. The Lipschitz constant for the corrected loss in the noise-free SD case is $L_Q = \frac{L}{|2\pi-1|}$, where L is the Lipschitz constant for ℓ. An estimation error bound for noise-free SD learning is:

Corollary 1. *For any $\delta > 0$, we have the following: with probability at least $1-\delta$,*

$$R_{\ell,\mathcal{Z}}(\hat{f}) \leq \min_{f\in\mathcal{F}} R_{\ell,\mathcal{Z}}(f) + \frac{4L}{|2\pi-1|}\mathcal{R}(\mathcal{F},2n) + 2\sqrt{\frac{\log(\frac{1}{\delta})}{4n}}. \tag{15}$$

$$R_{\mathcal{Z}}(\hat{f}) - R^* \leq \xi_\ell^{-1}\left(\min_{f\in\mathcal{F}} R_{l,\mathcal{Z}}(f) - \min_{f} R_{l,\mathcal{Z}}(f) + \frac{4L\mathcal{R}(\mathcal{F},2n)}{|2\pi-1|} + 2\sqrt{\frac{\log(\frac{1}{\delta})}{4n}}\right). \tag{16}$$

This is directly obtained from Theorem 1 by plugging in the value of L_Q.
Optimization: While we have a performance guarantee, efficient optimization is a concern especially because the corrected loss $\tilde{\ell}(\cdot,\cdot)$ may not be convex. We present a condition which will guarantee the corrected loss to be convex.

Theorem 2. *If $\ell(t, y)$ is convex and twice differentiable almost everywhere in t (for each y) and also satisfies:*

- $\forall t \in \mathbb{R}, \quad \ell'(t, y) = \ell'(t, -y)$, *where the differentiation is w.r.t. t.*
- $\operatorname{sign}(\alpha_1 - \beta_1) = \operatorname{sign}(\beta_2 - \alpha_2)$, *where $\alpha_1, \alpha_2, \beta_1, \beta_2$ are elements of T.*

then $\tilde{\ell}(t, y)$ is convex in t.

Proof of Theorem 2 is available in Appendix 5.

The first condition is satisfied by several common losses such as squared loss $\ell(t, y) = (1 - ty)^2$ and logistic loss $\ell(t, y) = \log(1 + \exp(-ty))$. The second condition depends on the noise rates and the class prior. We can simplify this for the pairing corruption noise model as:

$$\begin{aligned} \frac{1-2\rho_\mathrm{S}}{1-2\rho_\mathrm{D}} &\in \left[\frac{1-\pi}{\pi}, \frac{\pi}{1-\pi}\right] \quad \text{if} \quad \pi \quad \geq 0.5 \quad , \\ \frac{1-2\rho_\mathrm{S}}{1-2\rho_\mathrm{D}} &\in \left[\frac{\pi}{1-\pi}, \frac{1-\pi}{\pi}\right] \quad \text{if} \quad \pi \quad \leq 0.5 \quad . \end{aligned} \tag{17}$$

In the case of the labeling corruption noise model, this condition reduces to

$$\begin{aligned} \frac{1-2\rho_+}{1-2\rho_-} &\in \left[\frac{1-\tilde{\pi}}{\tilde{\pi}}, \frac{\tilde{\pi}}{1-\tilde{\pi}}\right] \quad \text{if} \quad \tilde{\pi} \quad \geq 0.5 \quad , \\ \frac{1-2\rho_+}{1-2\rho_-} &\in \left[\frac{\tilde{\pi}}{1-\tilde{\pi}}, \frac{1-\tilde{\pi}}{\tilde{\pi}}\right] \quad \text{if} \quad \tilde{\pi} \quad \leq 0.5 \quad . \end{aligned} \tag{18}$$

For all cases of noise-free or symmetric-noise ($\rho_S = \rho_D$ or $\rho_+ = \rho_-$) SD learning, any noise rates will satisfy this condition and thus, we can always perform efficient optimization. For cases where the above condition is not satisfied, i.e., $\tilde{l}$ is not guaranteed to be convex, this is often not a problem in practice since, neural networks optimized by stochastic gradient descent (our setup in Sec. 5) converges efficiently to a globally optimal solution, under certain conditions [6].

4 Weighted Classification Approach

Now we develop our second algorithm for dealing with noisy S and D data. One key issue that we investigate here is how the Bayes classifier learned from noisy SD data relates to the traditional Bayes classifier.

Lemma 2. *Denote the modified posterior under a SD noise model as $P(Q = 1|x) = \eta_Q(x)$ and $P(Y = 1|x) = \eta(x)$. Then the Bayes classifier under the noisy SD distribution $\tilde{f}^* = \arg\min_{f \in \mathcal{F}} \mathbb{E}_{(X,Q)\sim\mathcal{Z}_Q}[1_{\{\operatorname{sign}(f(X)) \neq Q\}}]$ is given by*

$$\tilde{f}^*(x) = \operatorname{sign}\left(\eta_Q(x) - \frac{1}{2}\right) = \operatorname{sign}\left(\eta(x) - \tau\right), \tag{19}$$

where, τ depends on the noise model, noise rates and π and is presented below.

For the pairwise corruption case, assuming $\pi \neq 0.5$,

$$\tau = \frac{\frac{1}{2} - [(1-\rho_{\mathrm{S}})(1-\pi) + \rho_{\mathrm{D}}\pi]}{(1-\rho_{\mathrm{S}}-\rho_{\mathrm{D}})(2\pi-1)}.$$

For the label corruption case, threshold τ *is:*

$$\begin{aligned}\tau &= \frac{\frac{1}{2} - \pi(\rho_+ + \rho_- - \rho_+\rho_-) - (1-\pi)(\rho_+^2 + (1-\rho_-)^2)}{(1-\rho_+-\rho_-)[\pi(1-2\rho_+) - (1-\pi)(1-2\rho_-)]}, \\ &= \frac{\frac{1}{2} - \pi(\rho_+ + \rho_- - \rho_+\rho_-) - (1-\pi)(\rho_+^2 + (1-\rho_-)^2)}{(1-\rho_+-\rho_-)(2\tilde{\pi}-1)}\end{aligned}$$

assuming $\tilde{\pi} \neq 0.5$ *where* $\tilde{\pi}$ *is defined in* (5).

These expressions can be derived by using (6) and (7) in (19) and the detailed proof of Lemma 2 is available in Appendix 6. They give us an important insight:

Remark 2. The Bayes classifier for noisy SD learning uses a different threshold from $\frac{1}{2}$ while the traditional Bayes classifier has $\eta(x)$ thresholded at $\frac{1}{2}$.

Towards designing an algorithm we note that we can also obtain this Bayes classifier by minimizing the weighted 0-1 risk defined as follows:

$$U_\alpha(t,y) = (1-\alpha)1_{\{y=1\}}1_{\{t\leq 0\}} + \alpha 1_{\{y=-1\}}1_{\{t>0\}}.$$

The following lemma from [22] is crucial in connecting the Bayes classifier threshold with the weight α in weighted 0-1 classification.

Lemma 3. *[22]: Denote the* U_α *risk under distribution* $\mathcal{Z}$ *as*

$$R_{\alpha,\mathcal{Z}}(f) = E_{(x,y)\sim\mathcal{Z}}[U_\alpha(f(x),y)].$$

Then $f_\alpha^*(x) = \mathrm{sign}(\eta(x)-\alpha)$ *minimizes* $R_{\alpha,\mathcal{Z}}(f)$.

We now show that there exists a choice of weight α such that the weighted risk under the noisy SD distribution is linearly related to the ordinary risk under distribution $\mathcal{Z}$.

Theorem 3. *There exist constants* α *and* A *and a function* $B(X)$ *that only depends on* X *but not on* f, *such that*

$$R_{\alpha,\mathcal{Z}_Q}(f) = AR_{\mathcal{Z}}(f) + E_X[B(X)].$$

For the pairing corruption case:

$$\alpha = \frac{1-\rho_{\mathrm{S}}+\rho_{\mathrm{D}}}{2}, \quad A = \frac{1-\rho_{\mathrm{S}}-\rho_{\mathrm{D}}}{2}(2\pi-1). \tag{20}$$

For the label corruption case:

$$\begin{aligned}\alpha &= \pi(1-\rho_+ + \rho_+^2 - \rho_+\rho_-) - \frac{1}{2}(1-\rho_+-\rho_-) + (1-\pi)(1-\rho_- + \rho_-^2 - \rho_+\rho_-), \\ A &= \frac{(1-\rho_+-\rho_-)}{2}[\pi(1-2\rho_+) - (1-\pi)(1-2\rho_-)].\end{aligned} \tag{21}$$

Proof of Theorem 3 is available in Appendix 7.

Remark 3. The α-weighted Bayes optimal classifier under the noisy SD distribution coincides with the Bayes classifier of the 0-1 loss under the standard distribution $\mathcal{Z}$.

$$\arg\min_{f} R_{\alpha*,\mathcal{Z}_Q}(f) = \arg\min_{f} R_{\mathcal{Z}}(f) = \text{sign}\left(\eta(x) - \frac{1}{2}\right).$$

Performance Bounds and Optimization: For the ease of optimization, we will use a surrogate loss instead of the 0-1 loss to perform weighted ERM. Any surrogate loss can be used as long as it can be decomposed as $\ell(t,Q) = 1_{\{Q=1\}}\ell_1(t) + 1_{\{Q=-1\}}\ell_{-1}(t)$, for partial losses ℓ_1, ℓ_{-1} of ℓ [3, 22]. The margin-based surrogate loss functions ℓ such that $\ell(t,Q) = 1_{\{Q=1\}}\phi(t) + 1_{\{Q=-1\}}\phi(-t)$ for some $\phi : \mathbb{R} \to \mathbb{R}_{\geq 0}$ [18] is expressible in this form. The commonly used surrogate losses such as the squared, hinge, and logistic losses are encompassed in the margin-based surrogate loss. We want to minimize the following empirical risk using the weighted surrogate loss l_α:

$$\min_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} \ell_\alpha(g(X_i, X_i'), Q_i). \tag{22}$$

Similar to (9) we consider the pointwise version of the empirical risk using the weighted surrogate loss l_α:

$$\min_{f \in \mathcal{F}} \frac{1}{2n} \sum_{i=1}^{2n} \ell_\alpha(f(X_i), Q_i), \tag{23}$$

and let $\hat{f}_\alpha$ denote the minimizer of (23).

$$\hat{f}_\alpha = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \ell_\alpha(f(x_i), Q_i). \tag{24}$$

We already discussed classification calibration in Section 3. For the following theorem we need the notion of α-classification calibrated losses, developed in [22], that extends [3] to the asymmetric classification setting where the misclassification costs are unequal for the two classes.

Theorem 4. *If ℓ_α is an α-weighted margin loss [22] of the form: $l_\alpha(t,Q) = (1-\alpha)1_{\{Q=1\}}\ell(t)+\alpha 1_{\{Q=-1\}}\ell(-t)$ and ℓ is convex, classification calibrated ($\ell'(0) < 0$ where the derivative is w.r.t. t) and L-Lipschitz, then for the choices of α and A in* (20),(21)*(assuming $\pi \neq 0.5$ or $\hat{\pi} \neq 0.5$ for the corresponding noise model), there exists a non-decreasing function ξ_{ℓ_α} with $\xi_{\ell_\alpha}(0) = 0$ such that the following bound holds with probability at least $1 - \delta$:*

$$R_{\mathcal{Z}}(\hat{f}_\alpha) - R^* \leq A^{-1}\xi_{\ell_\alpha}\left(\min_{f \in \mathcal{F}} R_{\alpha,\mathcal{Z}_Q}(f) - \min_{f} R_{\alpha,\mathcal{Z}_Q}(f) + 4L\mathcal{R}(\mathcal{F}, n) + 2\sqrt{\frac{\log(\frac{1}{\delta})}{2n}}\right), \tag{25}$$

where R^* denotes the the corresponding Bayes risk under $\mathcal{Z}$.

Note that using Corollary 4.1 from [22] we know l_α is α-classification calibrated. The right-side in (25) is finite because $A \neq 0$ whenever $\pi, \hat{\pi} \neq 0.5$. Proof of Theorem 4 is available in Appendix 8.

Remark 4. For a fixed Lipschitz constant L, as A decreases we get a weaker excess risk bound. For the pairing corruption noise model, its easy to see that as noise rates increase, A decreases. On the other hand, the relationship is more complicated for the labeling corruption noise model. When the noise is symmetric ($\rho_+ = \rho_- = \rho$), $A = \frac{(1-2\rho)^2(2\pi-1)}{2}$. In this case, again we observe as ρ increases, A decreases and we get a weaker bound.

Remark 5. When $\rho_{\mathrm{S}} = \rho_{\mathrm{D}}$ or $\rho_+ = \rho_-$, we see that the optimal Bayes classifier for the (noisy) SD learning problem is the same as the Bayes classifier for the standard class-labeled binary classification task under distribution $\mathcal{Z}$. In these settings, this result allows us to learn a classifier for standard class-labeled binary classification from (noisy) SD data simply by treating the similar and dissimilar classes as the positive and negative class for any chosen classifier.

Estimation of Prior and Noise Parameters: We briefly discuss the parameters (the class prior π and the noise rates ρ_{S} and ρ_{D} in the pairing corruption noise model and $\rho_\pm$ in the labeling corruption noise model) that we need to know or estimate to apply each method for each noise model.

(I) Loss Correction Approach: The noise rate parameters can be tuned by cross-validation on the noisy SD data. We also need to estimate the class prior to construct the loss correction matrix T, under both noise models. Let n_{S} be the number of similar pairs and n_{D} be the number of dissimilar pairs in the training dataset. The class prior π can be estimated from the following equations:

- For the pairing corruption noise model:

$$\frac{n_S}{n_D} \approx \frac{(1-\rho_{\mathrm{S}})(\pi^2 + (1-\pi)^2) + 2\rho_{\mathrm{D}}\pi(1-\pi)}{\rho_{\mathrm{S}}(\pi^2 + (1-\pi)^2) + 2(1-\rho_{\mathrm{D}})\pi(1-\pi)} \tag{26}$$

- For the labeling corruption noise model:

$$\frac{n_S}{n_D} \approx \frac{(1-\rho_+)(\pi\tilde{\pi} + (1-\pi)(1-\tilde{\pi})) + \rho_-(\pi(1-\tilde{\pi}) + \tilde{\pi}(1-\pi))}{\rho_+(\pi\tilde{\pi} + (1-\pi)(1-\tilde{\pi})) + (1-\rho_-)(\pi(1-\tilde{\pi}) + \tilde{\pi}(1-\pi))} \tag{27}$$

From each of the above equations we can obtain an estimate $\hat{\pi}$ of the class prior π . The above equations can be derived from (6) and (7) by marginalizing out X and using $\frac{n_S}{n_D} \approx \frac{P(Q=1)}{P(Q=-1)}$ (equality holds for the population).

(II) Weighted Classification Approach: The class prior only appears in the weight α in the labeling corruption model. In the pairing corruption model, knowledge of the class prior is not needed to calculate α. However, since we just have one parameter α for the optimization problem, in practice we can obtain α directly by cross-validation under both noise models. Note that if we are given the noise rates, in the pairing corruption noise model we can calculate the optimum α exactly but in the labeling corruption noise model we still get only an estimate of the optimum α since, $\hat{\pi} \approx \pi$.

5 Experiments

We empirically verify that the proposed algorithms are able to learn a classifier for the underlying distribution $\mathcal{Z}$ from only noisy similar and dissimilar training data. All experiments are repeated 3 times on random train-test splits of 75:25 and the average accuracies are shown. We conduct experiments on two noise models independently. In the learning phase, the noise parameters and the weight α is tuned by cross-validation for the Loss Correction Approach and the Weighted Classification Approach respectively, for both noise models, by searching in $[0, 0.5]$ in increments of 0.1. Evaluation is done on the standard class-labeled test dataset using standard classification accuracy (Eq.1) as evaluation metric which is averaged over the test datasets to reduce variance across the corruption in the training data. We use a multi-layer perceptron (MLP) with two hidden layers of 100 neurons, ReLU activation and a single logistic sigmoid output, as our model architecture for all experiments trained using the squared loss: $\ell(t, y) = (t - y)^2$. We use stochastic gradient descent with momentum of 0.9 with a mini-batch size of 32 and a learning rate of 0.001, for 500 epochs.

Synthetic Data: We use a non-separable benchmark "banana" dataset which has two dimensional attributes and two classes. We perform two kinds of experiments. In the first experiment, for a given noise model, for different settings of symmetric noise parameters ($\rho_{\mathrm{S}} = \rho_{\mathrm{D}}$ and $\rho_+ = \rho_-$) we plot the variation of standard test accuracy with the number of noisy SD pairs (n) sampled for training. For this experiment setting, we show the results for the weighted classification algorithm in Figure 2. Since the Bayes classifier under the symmetric noise is identical to that of noise-free case under both the noise models (see Remark 4), we see that the accuracy improves as we get a better approximation of the Bayes classifier as we have more SD data-points in training. Note that the number of original training points in the dataset is fixed—what changes is only the number of SD points we sample from them. In the second experiment, for each noise model, for a fixed n we show the gradual degradation of performance of the proposed algorithms (loss correction approach as well as the weighted clas-

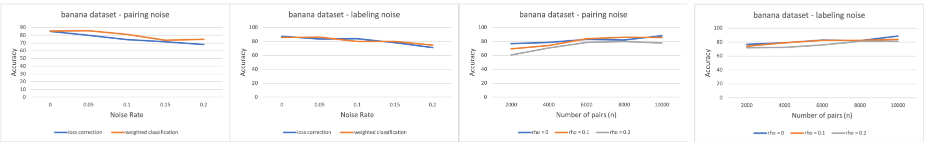

Fig. 2: The left two images depict the gradual decrease in classification accuracy of the learned classifier (from either algorithm) as the noise rate in the noisy SD training data increases, for each noise model. The right two images depict the increase in classification accuracy of the learned classifier (from the weighted classification method) as the number of noisy SD training samples increases, for different noise rates in each noise model. The accuracy achieved by training on standard P-N training data provided by the banana dataset is 90.8%.

Table 1: Pairing Noise : The best P-N column denotes the test accuracy after training on the standard class-labeled train dataset provided. d, π, N denote the feature dimension, class prior and the size of the entire class-labeled data respectively. Clean S-D denotes test accuracy after training on noise-free S-D data generated from the train dataset. T-Loss indicates the test accuracy after training on S-D data with the loss correction approach (by the matrix T) and SD-Loss denotes the non-corrected variant of [23]. Similarly, weighted and unweighted denotes the test accuracy after training on S-D data using weighted ERM and normal ERM respectively—note, they are identical for symmetric noise. KM denotes the KMeans baseline and KM-COP is the KMeans with constraints. Accuracies within 1% of the best in each row are bolded.

Dataset (d, π) N	best P-N	clean S-D	Noise Rates (ρ_S, ρ_D)	T-Loss	SD-Loss	Weighted	Unweight -ed	KM	KM- COP
DIABETES	77	77	$(0.2, 0.2)$	**76.57**	75	74.95	74.95	65.63	64.58
(8,0.35)			$(0.1, 0.2)$	74.95	75.52	**77.61**	76.04		65.10
768			$(0.3, 0.3)$	**75.52**	73.95	74.48	74.48		64.06
ADULT	83.09	83.03	$(0.2, 0.2)$	**82.49**	**82.22**	**82.35**	**82.35**	71.25	71.25
(106,0.24)			$(0.1, 0.2)$	77.8	76.26	**82.41**	75.92		71.25
48842			$(0.3, 0.3)$	**81.42**	80.26	**81.10**	**81.10**		53.14
CANCER	97.2	97.2	$(0.2, 0.2)$	**97.18**	**96.47**	95.78	95.78	88.7	92.95
(30,0.37)			$(0.1, 0.2)$	**97.18**	**97.18**	95.78	95.78		91.54
569			$(0.3, 0.3)$	**97.18**	95.07	95.78	95.78		92.25

sification approach) with increasing symmetric noise rates. These experiments confirm that higher noise hurts accuracy and more pairwise samples helps it.

Real World datasets We further conduct experiments on several benchmark datasets from the UCI classification tasks.[5] All tasks are binary classification tasks of varying dimensions, class priors, and sample sizes. We compare the performance of our proposed approaches against two kinds of baselines.

(A) Supervised Baselines: The state-of-the-art algorithm [23] for learning from pairwise similar-dissimilar data is used and this provides a strong baseline for the loss-correction approach. We also compare the performance of the weighted classification approach thresholded at $\frac{1}{2}$, i.e., under the noise-free assumption. While these baselines have been proved to perform very well in the noise-free scenario (both theoretically and empirically), here we investigate if they are robust to noisy annotations, for varying noise rates.

(B) Unsupervised Baselines: We also compare against unsupervised clustering and semi-supervised clustering based methods. For unsupervised clustering, pairwise information is ignored KMeans [16] is applied with $K = 2$ clusters, directly on the noisy SD datapoints and the obtained clusters are used to classify the test data. We also use constrained KMeans clustering [25], where we treat the SD pairs as *must-link* and *cannot-link* constraints to supervise the clustering of

[5] Available at https://archive.ics.uci.edu/ml/datasets.php.

Table 2: Labeling Noise: The setup is same as Table 1 but now we use the labeling corruption noise model to generate the noisy S-D data.

Dataset (d, π) N	best P-N	clean S-D	Noise Rates (ρ_+, ρ_-)	T-Loss	SD-Loss	Weight -ed	Unweight -ed	KM	KM- COP
IONOSPHERE			$(0.2, 0.2)$	**88.67**	85.23	86.4	86.4		71.59
(34,0.64)	90.91	90.91	$(0.1, 0.2)$	85.24	80.68	**90.91**	85.23	70.45	71.59
351			$(0.3, 0.3)$	87.5	80.7	**88.64**	**88.64**		71.59
SPAMBASE			$(0.2, 0.2)$	**87.56**	83.22	82.78	82.78		78.96
(57,0.39)	91.83	89.74	$(0.1, 0.2)$	83.74	84.15	**86.78**	85.56	78.08	79.13
4601			$(0.3, 0.3)$	**85.304**	75.65	78.44	78.44		78.61
MAGIC			$(0.2, 0.2)$	80.06	81.13	**82.21**	**82.21**		63.28
(10,0.65)	84.12	83.40	$(0.2, 0.1)$	73.27	73.61	**81.67**	79.70	59.09	66.03
19020			$(0.3, 0.3)$	**79.39**	78.42	**79.50**	**79.50**		62.34

the SD data pooled together. While constrained clustering is a strong baseline for pairwise learning [23], here we investigate if it is robust to noisy annotations.

In Tables 1 and 2, we show the performance of our proposed algorithms versus the baselines. We observe that for both noise models and for almost all noise rates, our proposed approaches *significantly* outperform the baselines. We also observe, that as the noise rates increase, performance degrades for all the methods. Further, we see that the noise-free SD performances match the best P-N performance which empirically verifies the optimal classifiers for learning from noise-free standard P-N and pairwise SD data coincide. The complete set of experiments along with additional details are provided in Appendix 9.

6 Conclusion and Future Work

In this paper we theoretically investigated a novel setting that is commonly encountered in several applications—learning from noisy pairwise labels and studied it under two distinct noise models. We showed the connections of this problem to standard class-labeled binary classification, proposed two algorithms and derived their performance bounds. We empirically showed that they outperform state-of-the-art supervised and unsupervised baselines and are able to handle severe noise corruption. For future work, it is worthwhile to investigate more complicated noise models such as instance-dependent noise [17] in this setting.

Acknowledgement: HB was supported by JSPS KAKENHI Grant Number 19J21094. MS was supported by JST AIP Acceleration Research Grant Number JPMJCR20U3, Japan. SD was an intern at RIKEN-AIP.

References

1. Bao, H., Niu, G., Sugiyama, M.: Classification from pairwise similarity and unlabeled data. In: International Conference on Machine Learning. pp. 461–470 (2018)

2. Bartlett, P.L., Bousquet, O., Mendelson, S., et al.: Local rademacher complexities. The Annals of Statistics **33**(4), 1497–1537 (2005)
3. Bartlett, P.L., Jordan, M.I., McAuliffe, J.D.: Convexity, classification, and risk bounds. Journal of the American Statistical Association **101**(473), 138–156 (2006)
4. Bartlett, P.L., Mendelson, S.: Rademacher and gaussian complexities: Risk bounds and structural results. Journal of Machine Learning Research **3**(Nov),(2002)
5. Basu, S., Davidson, I., Wagstaff, K.: Constrained clustering: Advances in algorithms, theory, and applications. CRC Press (2008)
6. Du, S.S., Zhai, X., Poczos, B., Singh, A.: Gradient descent provably optimizes over-parameterized neural networks. arXiv preprint arXiv:1810.02054 (2018)
7. Elkan, C.: The foundations of cost-sensitive learning. In: International joint conference on artificial intelligence. vol. 17, pp. 973–978. (2001)
8. Eric, B., Freitas, N.D., Ghosh, A.: Active preference learning with discrete choice data. In: Advances in neural information processing systems. pp. 409–416 (2008)
9. Fisher, R.J.: Social desirability bias and the validity of indirect questioning. Journal of consumer research **20**(2), 303–315 (1993)
10. Fürnkranz, J., Hüllermeier, E.: Preference learning. Springer (2010)
11. Gomes, R., Welinder, P., Krause, A., Perona, P.: Crowdclustering. In: NIPS (2011)
12. Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., Tsang, I., Sugiyama, M.: Co-teaching: Robust training of deep neural networks with extremely noisy labels. In: Advances in neural information processing systems. pp. 8527–8537 (2018)
13. Hsu, Y.C., Lv, Z., Schlosser, J., Odom, P., Kira, Z.: Multiclass classification without multiclass labels. International Conference on Learning Representations (2018)
14. Jamieson, K.G., Nowak, R.: Active ranking using pairwise comparisons. In: Advances in Neural Information Processing Systems. pp. 2240–2248 (2011)
15. Jiang, L., Zhou, Z., Leung, T., Li, L.J., Fei-Fei, L.: Mentornet: Regularizing very deep neural networks on corrupted labels. arXiv preprint arXiv:1712.05055 **4** (2017)
16. MacQueen, J., et al.: Some methods for classification and analysis of multivariate observations. In: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. vol. 1, pp. 281–297. Oakland, CA, USA (1967)
17. Menon, A.K., Van Rooyen, B., Natarajan, N.: Learning from binary labels with instance-dependent corruption. arXiv preprint arXiv:1605.00751 (2016)
18. Mohri, M., Rostamizadeh, A., Talwalkar, A.: Foundations of machine learning. MIT press (2018)
19. Natarajan, N., Dhillon, I.S., Ravikumar, P.K., Tewari, A.: Learning with noisy labels. In: Advances in neural information processing systems. pp. 1196–1204 (2013)
20. Patrini, G., Rozza, A., Krishna Menon, A., Nock, R., 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. (2017)
21. Saaty, T.L.: Decision making for leaders: the analytic hierarchy process for decisions in a complex world. RWS publications (1990)
22. Scott, C., et al.: Calibrated asymmetric surrogate losses. Electronic Journal of Statistics **6**, 958–992 (2012)
23. Shimada, T., Bao, H., Sato, I., Sugiyama, M.: Classification from pairwise similarities/dissimilarities and unlabeled data via empirical risk minimization. arXiv preprint arXiv:1904.11717 (2019)
24. Thurstone, L.L.:A law of comparative judgment. Psychological review **34**(4),(1927)
25. Wagstaff, K., Cardie, C., Rogers, S., Schrödl, S., et al.: Constrained k-means clustering with background knowledge. In: Icml. vol. 1, pp. 577–584 (2001)
26. Yi, J., Jin, R., Jain, A.K., Jain, S.: Crowdclustering with sparse pairwise labels: A matrix completion approach. In: HCOMP@ AAAI. Citeseer (2012)