Pairwise Supervision Can Provably Elicit a Decision Boundary

Han Bao[1,2,*]   **Takuya Shimada**[1,2,*,†]   **Liyuan Xu**[3]   **Issei Sato**[1]   **Masashi Sugiyama**[2,1]
[1]The University of Tokyo, Japan
[2]RIKEN AIP, Japan
[3]Gatsby Unit, UCL, UK
[*] Equal contribution (correspondence to Han Bao: tsutsumi@ms.k.u-tokyo.ac.jp)
[†] The author is now with Preferred Networks, Inc., Japan.

Abstract

Similarity learning is a general problem to elicit useful representations by predicting the relationship between a pair of patterns. This problem is related to various important pre-processing tasks such as metric learning, kernel learning, and contrastive learning. A classifier built upon the representations is expected to perform well in downstream classification; however, little theory has been given in literature so far and thereby the relationship between similarity and classification has remained elusive. Therefore, we tackle a fundamental question: can similarity information provably leads a model to perform well in downstream classification? In this paper, we reveal that a product-type formulation of similarity learning is strongly related to an objective of binary classification. We further show that these two different problems are explicitly connected by an excess risk bound. Consequently, our results elucidate that similarity learning is capable of solving binary classification by directly eliciting a decision boundary.

1 Introduction

Similarity learning is a learning paradigm (Kulis, 2013) that builds a pairwise model to predict whether given paired patterns are similar or dissimilar in the classes that they belong to. We call such a pair of patterns *pairwise supervision*, in contrast to ordinary *pointwise*

supervision which binds a class label to a single input pattern. Pairwise supervision is commonly available in many domains such as geographical analysis (Wagstaff et al., 2001), chemical experiment (Eisenberg et al., 2000), click-through feedback (Davis et al., 2007), computer vision (Yan et al., 2006; Wang and Gupta, 2015), natural language processing (Mikolov et al., 2013), and crowdsourcing (Gomes et al., 2012). Notably, feature representations can be constructed from pairwise supervision when it is not straightforward to define meaningful features (Chen et al., 2009; Wang et al., 2009; Kar and Jain, 2011). This is one of the reasons why similarity learning has been studied extensively—including metric learning (Xing et al., 2003; Bilenko et al., 2004; Davis et al., 2007; Weinberger and Saul, 2009; Bellet et al., 2012; Niu et al., 2014), kernel learning (Cristianini et al., 2002; Bach et al., 2004; Lanckriet et al., 2004; Li and Liu, 2009; Cortes et al., 2010), and $(\varepsilon, \gamma, \tau)$-good similarity (Balcan et al., 2008; Wang et al., 2009; Kar and Jain, 2011; Bellet et al., 2012), with different notion of similarity and models. In recent studies, a similarity model is trained that aligns with pairwise supervision to capture inherent structures of data (Bellet et al., 2012; Mikolov et al., 2013; Niu et al., 2014; Logeswaran and Lee, 2018; Saunshi et al., 2019). The learned similarity model is expected to help downstream tasks. Correspondingly, it has been widely used for various downstream tasks such as classification (Cristianini et al., 2002; Balcan et al., 2008; Hsu et al., 2019; Saunshi et al., 2019; Nozawa et al., 2020), clustering (Bromley et al., 1994; Xing et al., 2003; Davis et al., 2007; Weinberger and Saul, 2009), model selection (Lanckriet et al., 2004), and one-shot learning (Koch et al., 2015).

The early theoretical research provided error bounds on classification based on similarity-based features by assuming that a given similarity metric is $(\varepsilon, \gamma, \tau)$-good (Balcan et al., 2008; Wang et al., 2009) (see related work for the details). Recently, it has been attempted to investigate the relationship between learned similarity models and downstream classification, in or-

der to deal with more flexible data structures. Bellet et al. (2012) proved that features based on a learned metric are linearly separable under the framework of $(\varepsilon, \gamma, \tau)$-good similarity. Saunshi et al. (2019) analyzed how features learned in contrastive learning are meaningful in downstream classification. These results boil down to two-step learners, which first solve similarity learning then train classifiers. However, the latter step often requires as many samples as the former step because the feature space constructed from the similarity function often becomes high-dimensional (see Bellet et al. (2012, Theorem 1) for details).

In this work, we pose a question on what formulation of similarity learning is directly connected to downstream classification and reveal that similarity learning with a model $f(\boldsymbol{x}) \cdot f(\boldsymbol{x}')$ has a monotonic relationship to binary classification with a classifier $f(\boldsymbol{x})$. This interrelation provides a new insight that *a binary decision boundary can essentially be obtained with only pairwise supervision* up to label permutation. The post-process determining correct class assignments once classes are separated becomes less label-demanding than the previous formulations (Bellet et al., 2012; Saunshi et al., 2019). While it is rather straightforward to use pointwise supervision to determine correct class assignments, we further found that pairwise supervision is sufficient for this purpose given that we know the majority class. Our results are notable in that: (i) we unravel that similarity learning enables us to implicitly elicit a binary decision boundary without any explicit training of classifiers, and (ii) the post-process is less costly in terms of pointwise supervision. Specifically, we will see: similarity learning is tied to the binary classification error up to label permutation (Section 3.1). The post-process to determine correct class assignments is discussed (Section 3.2). As a by-product, we come across a training method of binary classifiers with only pairwise supervision (Section 3.3). A finite-sample excess risk bound is established to connect similarity learning to binary classification (Section 4). This theoretical finding is numerically demonstrated (Section 5).

Remark 1 (Multi-class case)**.** *Despite that our main result (Theorem 1) is limited to the binary case, we can apply our training method (described in Section 3.3) in the multi-class case by the one-vs-rest approach: given C classes, our training method can provide an one-vs-rest classifier for class $i \in [C]$ from pairwise supervision treating $[C] \setminus \{i\}$ as a single class. Kar and Jain (2011) took the same one-vs-rest approach to first automatically construct feature representations from similarity information, and then train the one-vs-rest multi-class classifier with pointwise labels. This approach is valuable in the domains where samples are not immediately accessible in a Euclidean space yet sophisticated distance metrics have been developed, such as graphs, sequences,* *and logics. See Ontañón (2020) and references therein for many examples. Nevertheless, we do not have any theoretical grounding of this approach so far.*

Related work. We review several variants of similarity models used in existing literature.

(A) Reliable similarity. This line of work regards two data as similar if the associated labels are the same. Our study and Bellet et al. (2012) belong to this category. Zhang and Yan (2007) proposed a method to decompose a model predicting pairwise labels into pointwise classifiers and analyzed the consistency of the model parameters. Hsu et al. (2019) have recently extended to the multi-class setup without theoretical justification yet. In parallel, other research solved classification with pairwise supervision by minimizing unbiased classification risk estimators (Bao et al., 2018; Shimada et al., 2021; Cui et al., 2020). Their approaches are blessed with generalization error bounds, while their performance deteriorates when the class-prior probability is close to uniform. Note that even the reliable similarity can handle mild noise in pairwise supervision (Remark 3). Recently, Tosh et al. (2021) revealed that pairwise supervision is sufficient to recover topic distributions under certain topic modeling assumptions.

(B) Noisy similarity. In this category, it is assumed that pairwise supervision aligns to the classes potentially with explicit noise. For example, negative samples in contrastive learning are usually drawn from the marginal distribution, hence they could be false negatives. Recently, Chuang et al. (2020) used techniques of unbiased risk estimators to improve the quality of negatives. Further, contrastive learning often assumes that similar pairs share the same latent category, which can be different from downstream supervised classes. Since contrastive learning is usually unsupervised, the supervised classes could be a subset or coarse-grained set of latent categories (Saunshi et al., 2019). Other research modeled annotation errors in pairwise supervision (Wu et al., 2020; Dan et al., 2021).

(C) Relaxation of positive-definite kernels. Balcan et al. (2008) introduced $(\varepsilon, \gamma, \tau)$-good similarity to relax positive-definiteness of kernel functions, which supposes that a good similarity function is useful for downstream linear classification. Much research in this framework has been interested in classification given features based on this weak similarity and derived classification error bounds (Balcan et al., 2008; Wang et al., 2009; Kar and Jain, 2011). Note that Bellet et al. (2012) assume reliable similarity as supervision and trains a similarity model while they train the model based on $(\varepsilon, \gamma, \tau)$-good similarity.

2 Problem setup

Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a d-dimensional pattern space, $\mathcal{Y} = \{\pm 1\}$ be the label space, and $p(\boldsymbol{x}, y)$ be the density of an underlying distribution over $\mathcal{X} \times \mathcal{Y}$. Denote the positive (negative, resp.) class prior by $\pi_+ := p(y = +1)$ ($\pi_- := p(y = -1)$, resp.). Let $\mathrm{sign}(\alpha) = 1$ for $\alpha > 0$ and -1 otherwise. $O_p(\cdot)$ denotes the order in probability.

Binary classification. The goal of binary classification is to classify unseen patterns into two classes. It can be formulated as a problem to find a classifier $h : \mathcal{X} \to \mathcal{Y}$ that minimizes

$$R_{\mathrm{point}}(h) := \mathbb{E}_{(X,Y) \sim p(\boldsymbol{x}, y)} [\mathbb{1}\{h(X) \neq Y\}], \quad (1)$$

where $\mathbb{1}\{\cdot\}$ is the indicator function and $\mathbb{E}_{(X,Y) \sim p(\boldsymbol{x}, y)}[\cdot]$ denotes the expectation with respect to $p(\boldsymbol{x}, y)$. Typically, we specify a hypothesis class $\mathcal{H}$ beforehand and find a minimizer h^* of R_{point} in it: $h^* \in \arg\min_{h \in \mathcal{H}} R_{\mathrm{point}}(h)$. The empirical mean of R_{point} is computed with finite samples.

Similarity learning. There are a variety of formulations of similarity learning such as (i) predicting whether a pair of patterns belong to the same class (Zhang and Yan, 2007; Bellet et al., 2012; Hsu et al., 2019), (ii) learning a metric that regards a similar pair of patterns closer (Bilenko et al., 2004; Davis et al., 2007; Niu et al., 2014; Vogel et al., 2018), and (iii) learning a metric/representation that represents a similar pair more closer than background samples (Wang et al., 2009; Kar and Jain, 2011; Saunshi et al., 2019). Specifically, we focus on the formulation (i) in the binary setup, which has a direct connection to classification (see Section 3). Hereafter, we suppose that a pair of $(\boldsymbol{x}, y)$ and $(\boldsymbol{x}', y')$ is independent of each other. Let $\eta_{\pm 1}(\boldsymbol{x}) := p(Y = \pm 1 | X = \boldsymbol{x})$. Assume that $X = \boldsymbol{x}$ and $X' = \boldsymbol{x}'$ are observed first and pairwise supervision T is drawn from

$$p(T = YY' | \boldsymbol{x}, \boldsymbol{x}')$$
$$= \begin{cases} \eta_{+1}(\boldsymbol{x}) \eta_{+1}(\boldsymbol{x}') + \eta_{-1}(\boldsymbol{x}) \eta_{-1}(\boldsymbol{x}') & \text{if } YY' = +1, \\ \eta_{+1}(\boldsymbol{x}) \eta_{-1}(\boldsymbol{x}') + \eta_{-1}(\boldsymbol{x}) \eta_{+1}(\boldsymbol{x}') & \text{if } YY' = -1. \end{cases}$$

The product YY' indicates whether Y and Y' are the same/similar ($+1$) or not/dissimilar (-1). Then, we are interested in the minimizer of the following classification error

$$R_{\mathrm{pair}}(h) := \mathbb{E}_{\substack{X, X' \sim p(\boldsymbol{x}) \\ T \sim p(T = YY' | \boldsymbol{x}, \boldsymbol{x}')}} [\mathbb{1}\{h(X) \cdot h(X') \neq T\}]. \quad (2)$$

Here, the model $h(\boldsymbol{x}) \cdot h(\boldsymbol{x}')$ is regarded as a similarity model so that we predict label agreement. We call R_{point} the *pointwise classification error* and R_{pair} the *pairwise classification error*. The empirical mean of R_{pair} is computed with a finite number of triplets $(\boldsymbol{x}, \boldsymbol{x}', yy')$. We will discuss several benefits of the formulation (2) in Section 3.4.

Remark 2 (Similarity as features). *Similarity-based features are often used in domains where Euclidean features are unavailable (Chen et al., 2009; Wang et al., 2009). Under such a case, similarity-based features may be treated as $\boldsymbol{x}$ instead: given a number of "landmark" points $\{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_l\}$, a similarity function K defines similarity-based features $[K(\boldsymbol{x}, \boldsymbol{z}_1), \ldots, K(\boldsymbol{x}, \boldsymbol{z}_l)]^\top$ for an input $\boldsymbol{x}$. Our formulation assumes that $\boldsymbol{x}$ is available for simplicity but can be replaced with similarity-based features.*

Remark 3 (T is not a hard similarity label). *Even if $Y = Y' = +1$ (similar) with high probability, we could observe $T = -1$ (dissimilar) with some probability. Assume $\eta_{+1}(\boldsymbol{x}), \eta_{+1}(\boldsymbol{x}') \in (\frac{1}{2}, 1)$. Then, the flipping rate $p(T = -1 | \boldsymbol{x}, \boldsymbol{x}')$ lies in $(0, \frac{1}{2})$. This means that observed pairwise supervision could be flipped stochastically under our similarity model. We expect that this generality is useful to handle annotation noise in pairwise supervision.*

3 Learning a binary classifier with pairwise supervision

We draw a connection between the specific formulation of similarity learning (2) and binary classification (Theorem 1). This linkage enables us to train a pointwise binary classifier with pairwise supervision (Section 3.3). All proofs hereafter are deferred to Appendix A.

3.1 Connection between similarity learning and classification

We first introduce a performance metric for binary classification called the *clustering error* that quantifies the discriminative power of a classifier up to label permutation:[1]

$$R_{\mathrm{clus}}(h) := \min\{R_{\mathrm{point}}(h), R_{\mathrm{point}}(-h)\}. \quad (3)$$

Here, R_{clus} is used as an evaluator of binary classifiers, though usually used for the evaluation of clustering methods (Fahad et al., 2014). The clustering error differs from R_{point} in that it dismisses the difference between $+h$ and $-h$, yet a binary decision boundary is still evaluated properly. The clustering error R_{clus} can be tied to the pairwise classification error R_{pair} as follows, which is our primary result.

---

[1] $1 - R_{\mathrm{clus}}$ is known as clustering accuracy (Fahad et al., 2014). The number of clusters is confined to two for our purpose.

Theorem 1. *For any classifier $h : \mathcal{X} \to \mathcal{Y}$, $0 \le R_{\mathrm{pair}}(h) \le \frac{1}{2}$, and*

$$R_{\mathrm{clus}}(h) = \frac{1}{2} - \frac{\sqrt{1 - 2R_{\mathrm{pair}}(h)}}{2}. \qquad (4)$$

An immediate corollary is the monotonic relationship $R_{\mathrm{clus}}(h_1) < R_{\mathrm{clus}}(h_2) \iff R_{\mathrm{pair}}(h_1) < R_{\mathrm{pair}}(h_2)$ for any h_1 and h_2. Hence, the minimization of R_{pair} amounts to the minimization of R_{clus}, constituting a decision boundary. That is, *similarity learning can essentially discover a binary decision boundary.* While similarity learning has previously been connected to downstream classification via intermediate feature spaces (Bellet et al., 2012; Saunshi et al., 2019; Nozawa et al., 2020), our result is the first to explicate that similarity learning is directly related to constructing a decision boundary.

Surrogate risk minimization. Here, we discuss surrogate losses for similarity learning. We define a hypothesis class by $\mathcal{H} = \{\mathrm{sign} \circ f \mid f \in \mathcal{F}\}$, where $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$ is a specified class of prediction functions and $\mathrm{sign} \circ f(\cdot) := \mathrm{sign}(f(\cdot))$. Theorem 1 suggests that we may minimize R_{clus} by minimizing R_{pair} instead. As in the standard binary classification case, the indicator function appearing in R_{pair} is replaced with a surrogate loss $\ell : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}_{\ge 0}$ since it is intractable to minimize a discrete objective (Bartlett et al., 2006). Eventually, the pairwise surrogate risk

$$R_{\mathrm{pair}}^{\ell}(f) := \underset{X,X',T}{\mathbb{E}} [\ell(f(X)f(X'), T)] \qquad (5)$$

is minimized. If ℓ is *classification-calibrated* (Bartlett et al., 2006), the minimization of R_{pair}^{ℓ} is expected to lead to minimizing R_{pair} as well.[2] This will be justified by Lemma 1 in Section 4.

As we will discuss in Section 3.4, the formulation (5) can be related to several existing formulations in similarity learning in terms of the surrogate loss.

3.2 Determination of correct sign of classifiers

For a given hypothesis h, we are now interested in its sign, i.e., $+h$ or $-h$, leading to a smaller pointwise classification error. We refer to this step as class assignment. The optimal class assignment is denoted by $s^* := \arg\min_{s \in \{\pm 1\}} R_{\mathrm{point}}(s \cdot h)$. We consider two scenarios. Under both, class assignment is much cheaper in supervision than training the post-hoc linear separators.

---

[2] If a surrogate loss ℓ is classification-calibrated, the minimization of the surrogate classification risk leads to minimizing the target classification error R_{point}. The precise definition can be found in Bartlett et al. (2006). Typical loss functions such as the logistic and hinge losses are classification-calibrated.

Class assignment with pointwise supervision. If pointwise supervision is available, we can determine the class assignment by minimizing the pointwise classification error R_{point} computed with the additional data. This procedure admits the exponentially small sample complexity (Zhang and Yan, 2007).

Class assignment without pointwise supervision. Here, we further ask if it is possible to obtain the correct class assignment *without* any class labels. Surprisingly, we find that this is possible if the positive and negative proportions are not equal and we know *which class is the majority*. Based on the equivalent expression of R_{point} (Shimada et al., 2021), this finding is formally stated in the following theorem.

Theorem 2. *Assume that the class prior $\pi_+ \ne \frac{1}{2}$. Then, the optimal class assignment s^* can be represented as $s^* = \mathrm{sign}(2\pi_+ - 1) \cdot \mathrm{sign}(1 - 2Q(h))$, where*

$$Q(h) := \underset{X,X',T}{\mathbb{E}} \left[\frac{\mathbb{1}\{h(X) \ne T\} + \mathbb{1}\{h(X') \ne T\}}{2} \right].$$

We approximate Q with a finite number of pairs. As we will see in Lemma 3 in Section 4, the class assignment error is exponentially small in the number of pairs.

Remark 4 (Necessity of $Q(h)$). *If we know which class is the majority, class assignment may look possible at a glance by simply looking at the average of $h(\boldsymbol{x})$ with unlabeled validation data, instead of Theorem 2. Unfortunately, this does not always succeed even asymptotically (discussed in Appendix B).*

3.3 Learning a binary classifier with only pairwise supervision is possible

As a by-product of Theorems 1 and 2, the following two-stage method can train a pointwise classifier with only pairwise supervision. Assume that the class prior is not $\frac{1}{2}$ and the majority class is known. Let $\mathfrak{D}_{\mathrm{train}} := \{(\boldsymbol{x}_i, \boldsymbol{x}'_i, \tau_i)\}_{i=1}^{n_{\mathrm{pair}}}$ be a training set, where $\tau_i := y_i y'_i$ and $(\boldsymbol{x}_i, y_i)$ and $(\boldsymbol{x}'_i, y'_i)$ are i.i.d. samples following $p(\boldsymbol{x}, y)$. We randomly divide n_{pair} pairs in $\mathfrak{D}_{\mathrm{train}}$ into two sets $\mathfrak{D}_1$ and $\mathfrak{D}_2$, where $|\mathfrak{D}_1| = m_1$ and $|\mathfrak{D}_2| = m_2$ satisfying $m_1 + m_2 = n_{\mathrm{pair}}$.[3]

In Step 1, we obtain a minimizer of the empirical pairwise classification risk with $\mathfrak{D}_1$:

$$\widehat{f} := \arg\min_{f \in \mathcal{F}} \widehat{R}_{\mathrm{pair}}^{\ell}(f), \qquad (6)$$

where $\widehat{R}_{\mathrm{pair}}^{\ell}$ is the sample mean of R_{pair}^{ℓ} with $\mathfrak{D}_1$. In

---

[3] The independent two sets are necessary otherwise errors of Steps 1 and 2 correlate, which leads to overfitting. Technically, they are required because Theorem 3 relies on the union bound.

Step 2, we assign classes with $\text{sign} \circ \widehat{f}$ and $\mathfrak{D}_2$:

$$\widehat{s} := \text{sign}(2\pi_+ - 1) \cdot \text{sign}(1 - 2\widehat{Q}(\text{sign} \circ \widehat{f})), \quad (7)$$

where $\widehat{Q}$ is the sample mean of Q with $\mathfrak{D}_2$. After all, $\widehat{s} \cdot \text{sign} \circ \widehat{f}$ is a desideratum. If class assignment is not necessary and just separating test patterns into two disjoint groups is the goal, we may simply set $m_1 = n_{\text{pair}}$ and omit Step 2 of finding $\widehat{s}$.

Remark 5 (Case of $\pi = \frac{1}{2}$). *With only pairwise supervision, class assignment is hopeless because both classes are essentially symmetric, while it is still possible to draw a decision boundary. Class assignment with pointwise supervision is still possible.*

3.4 Benefits of our formulation over existing similarity learning

We reiterate that similarity learning in our formulation directly elicits a boundary without the post-process in contrast with Bellet et al. (2012)—their method needs to train a classifier built on top of the learned similarity metric in the post-process, which incurs additional sample complexity $O_p(m^{-1/2})$. Table 1 provides an overview of the comparison with related work. We remark that the sample complexity of SLLC is transformed into the complexity in terms of paired data (Step 1) from the original complexity in pointwise data (Bellet et al., 2012, Theorem 3).[4] While our Step 1 is worse than SD, our formulation is valid even when $\pi_+ = \frac{1}{2}$ with pointwise supervision. Subsequently, we discuss the other perspectives of our formulation.

Generalization in terms of surrogate losses. Several existing formulations can be related to our formulation (5). Kernel alignment (Cristianini et al., 2002) learns a kernel K approximating a similarity matrix K^* of labels by maximizing the cosine similarity $\frac{\langle K, K^* \rangle}{\sqrt{\|K\| \cdot \|K^*\|}}$, where $\langle K, K^* \rangle$ is the Frobenius inner product of the Gram matrices. If the product $f(\boldsymbol{x}) \cdot f(\boldsymbol{x}')$ is used as a kernel, kernel alignment is equivalent (up to the normalization factor $\sqrt{\|K\|\|K^*\|}$) to minimizing Eq. (5) with the linear loss $\ell_{\text{lin}}(z, t) := -zt$. On the other hand, metric learning based on $(\varepsilon, \gamma, \tau)$-good similarity (Balcan et al., 2008) regards a similarity function inducing a good linear separator as a good similarity. Here, the linear separability is defined via the hinge loss $\ell_{\text{hinge}}(z, t) := [1 - zt]_+$. Bellet et al. (2012) formulated learning a bilinear similarity $\boldsymbol{x}^\top A \boldsymbol{x}'$ by minimizing the hinge loss, which is equivalent to the minimization of

Eq. (5) with ℓ_{hinge} and the choice $A = \boldsymbol{w}\boldsymbol{w}^\top$ such that $f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x}$. In other words, we posit the rank-1 similarity model in order to have Theorem 1. In addition to these examples, the InfoNCE loss used in recent contrastive learning (van den Oord et al., 2018) can be regarded as the (multi-sample counterpart of) logistic loss $\ell_{\text{log}}(z, t) := \log(1 + e^{-zt})$.

Thanks to this generalization, subsequent analysis systematically connects these existing formulations to downstream classification under the model assumption.

Explicit relation to classification. Hsu et al. (2019) formulated similarity learning in a slightly different way, as maximum likelihood estimation of the pairwise label $S_\tau := \frac{\tau+1}{2}$:[5]

$$\min_{f \in \mathcal{F}} \frac{1}{m_1} \sum_{(\boldsymbol{x}, \boldsymbol{x}', \tau) \in \mathfrak{D}_1} -S_\tau \log(\tilde{q}(f(\boldsymbol{x}), f(\boldsymbol{x}'))) \qquad (8)$$
$$- (1 - S_\tau) \log(1 - \tilde{q}(f(\boldsymbol{x}), f(\boldsymbol{x}'))),$$

where $\tilde{q}(z, z') := \begin{bmatrix} q(z) \\ 1-q(z) \end{bmatrix}^\top \begin{bmatrix} q(z') \\ 1-q(z') \end{bmatrix}$ is the inner product of two binary probability vectors, and $q(z) := (1 + \exp(-z))^{-1}$ denotes the (inverse) logit link. On the other hand, our formulation (6) with the logistic loss $\ell_{\text{log}}(z, t) = -S_t \log(q(z)) - (1 - S_t) \log(1 - q(z))$ is

$$\min_{f \in \mathcal{F}} \frac{1}{m_1} \sum_{(\boldsymbol{x}, \boldsymbol{x}', \tau) \in \mathfrak{D}_1} -S_\tau \log(q(f(\boldsymbol{x}) \cdot f(\boldsymbol{x}'))) \qquad (9)$$
$$- (1 - S_\tau) \log(1 - q(f(\boldsymbol{x}) \cdot f(\boldsymbol{x}'))).$$

In the formulation (8), similarity is defined by the inner product of class probabilities, while it is defined by the inner product of f in the formulation (9). The latter definition is often called the *inner product similarity* (IPS) model (Okuno and Shimodaira, 2020).[6] While both are valid similarity learning methods, the IPS model (9) has several benefits: one can choose arbitrary loss functions,[7] and besides, the pairwise classification risk minimization (6) admits an excess risk bound (Lemma 1 in Section 4). For this reason, we call our formulation *CIPS (Classifier with Inner Product Similarity)* from now on.

---

[4] Given m pointwise data, $O(m^2)$ pairs can be generated and thereby the sample complexity is transformed. Strictly speaking, the generated $O(m^2)$ points are not independent of each other. Nevertheless, the convergence rate would remain the same by using the error bound with interdependent data (Usunier et al., 2005).

[5] The multi-class formulation in Hsu et al. (2019) was simplified in binary classification here for comparison.

[6] The IPS model originally defined similarity between two vector data representations, hence is called *inner* product similarity. Yet, the IPS model is applied on one-dimensional prediction $f(\boldsymbol{x})$ in our context. The IPS model has been used in several domains (Tang et al., 2015; Logeswaran and Lee, 2018; Saunshi et al., 2019; Okuno and Shimodaira, 2020).

[7] The formulation (8) can be extended from maximum likelihood estimation by using an arbitrary proper scoring rules (Gneiting and Raftery, 2007), but non-proper losses such as the hinge loss cannot be used.

Table 1: Comparison of closely related methods to train classifiers with pairwise supervision. They assume the availability of reliable similarity (see Section 1). The column "$\pi_+ = \frac{1}{2}$" shows whether the formulation is valid under $\pi_+ = \frac{1}{2}$. In *sample complexity*, m denotes the number of paired data in Step 1, and either paired or pointwise data in Step 2. The sample complexity analysis of Step 1 is with respect to either pointwise classification or clustering error. To make the comparison proper, we assume that the hinge loss is used and eventually ψ-transform is $\psi(u) = u$. This is detailed in Section 4 (Discussion).

| | | Sample complexity of | | |
	$\pi_+ = \frac{1}{2}$	Similarity learning (Step 1)	Post-process (Step 2)	Comment
CIPS (Ours)	✓	$O_p(m^{-\frac{1}{4}})$ (Lemma 2 in §4)	$O_p(e^{-m})$ (Lemma 3 in §4)	Step 2 is class assignment.
OVPC (Zhang and Yan, 2007)	✓	(N/A)	$O_p(e^{-m})$	Step 2 is class assignment. Step 1 was shown to be consistent but complexity is not known.
SLLC (Bellet et al., 2012)	✓	$O_p(m^{-\frac{1}{4}})$	$O_p(m^{-\frac{1}{2}})$	Step 2 is SVM training.
MCL (Hsu et al., 2019)	✓	(N/A)	(N/A)	Inner product of classifiers is fitted in Step 1. Sample complexities have yet to be known.
SD (Shimada et al., 2021)	–	$O_p(m^{-\frac{1}{2}})$	(unnecessary)	Step 1 trains classifiers directly.

4 Excess risk and sample complexity analysis

In this section, we provide the sample complexity analyses of CIPS in Table 1 and the excess risk bound to claim that CIPS does solve binary classification.

Let $\widehat{f}$ and $\widehat{s}$ be the solutions of Eqs. (6) and (7), respectively. The target excess risk for similarity learning is denoted by

$$\mathrm{Er}_{\mathrm{point}}(\widehat{s} \cdot \mathrm{sign} \circ \widehat{f}) := R_{\mathrm{point}}(\widehat{s} \cdot \mathrm{sign} \circ \widehat{f}) - R_{\mathrm{point}}^*,$$

where $R_{\mathrm{point}}^* := \inf_f R_{\mathrm{point}}(\mathrm{sign} \circ f)$, and $\inf_f$ indicates the infimum over all measurable functions. In addition, we introduce notation for the other excess risks:

$$\mathrm{Er}_{\mathrm{clus}}(\mathrm{sign} \circ f) := R_{\mathrm{clus}}(\mathrm{sign} \circ f) - R_{\mathrm{clus}}^*,$$
$$\mathrm{Er}_{\mathrm{pair}}(\mathrm{sign} \circ f) := R_{\mathrm{pair}}(\mathrm{sign} \circ f) - R_{\mathrm{pair}}^*,$$
$$\mathrm{Er}_{\mathrm{pair}}^\ell(f) := R_{\mathrm{pair}}^\ell(f) - R_{\mathrm{pair}}^{\ell,*},$$

where $R_{\mathrm{clus}}^* := \inf_f R_{\mathrm{clus}}(\mathrm{sign} \circ f)$. R_{pair}^* and $R_{\mathrm{pair}}^{\ell,*}$ are defined as the infima over all measurable functions similarly. To derive the excess risk bound on $\mathrm{Er}_{\mathrm{point}}(\widehat{s}\cdot\mathrm{sign} \circ \widehat{f})$, we need to handle errors of clustering error minimization and class assignment independently, which will be shown in Lemmas 2 and 3, respectively. An important insight to combine two errors is that if the class assignment is successful, $\mathrm{Er}_{\mathrm{point}}(\widehat{s} \cdot \mathrm{sign} \circ \widehat{f})$ is equivalent to the excess risk of clustering error minimization. That is to say,

$$\widehat{s} = \arg\min_{s \in \{\pm 1\}} R_{\mathrm{point}}(s \cdot \mathrm{sign} \circ \widehat{f})$$
$$\implies \mathrm{Er}_{\mathrm{point}}(\widehat{s} \cdot \mathrm{sign} \circ \widehat{f}) = \mathrm{Er}_{\mathrm{clus}}(\mathrm{sign} \circ \widehat{f}). \quad (10)$$

To bound $\mathrm{Er}_{\mathrm{clus}}(\mathrm{sign} \circ \widehat{f})$, we use the Rademacher complexity (Bartlett and Mendelson, 2002) specifically defined on the class $\{(\boldsymbol{x}, \boldsymbol{x}') \mapsto f(\boldsymbol{x}) \cdot f(\boldsymbol{x}') \mid f \in \mathcal{F}\}$

$$\mathfrak{R}_m(\mathcal{F}) := \mathbb{E}_{X_i, X_i'} \left[\sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^m \sigma_i f(X_i) \cdot f(X_i') \right],$$

where $\{\sigma_i\}_{i=1}^m$ are the Rademacher variables. Before obtaining an excess risk bound of R_{clus}, we need to bridge the excess risk $\mathrm{Er}_{\mathrm{pair}}$ and the surrogate $\mathrm{Er}_{\mathrm{pair}}^\ell$.

Lemma 1. *For a classification-calibrated loss ℓ (Bartlett et al., 2006), then there exists a convex, non-decreasing, and invertible $\psi : [0,1] \to [0, +\infty)$ such that for any sequence (u_i) in $[0,1]$,*

$$\psi(u_i) \to 0 \quad \text{if and only if} \quad u_i \to 0$$

and for any measurable function f and probability distribution on $\mathcal{X} \times \mathcal{Y}$,

$$\psi(\mathrm{Er}_{\mathrm{pair}}(\mathrm{sign} \circ f)) \le \mathrm{Er}_{\mathrm{pair}}^\ell(f).$$

Although the similar result to Lemma 1 has already been known for R_{point} (Bartlett et al., 2006, Theorem 1), the proof for R_{pair} requires special care to treat the product of prediction functions properly.

Then, the excess risk bound for R_{clus} is derived based on Lemma 1 and the uniform bound.

Lemma 2. *Let $f^* \in \mathcal{F}$ be a minimizer of R_{pair}^ℓ, and $\widehat{f} \in \mathcal{F}$ be a minimizer of $\widehat{R}_{\mathrm{pair}}^\ell$ defined in Eq. (6). Assume that $\ell(\cdot, \pm 1)$ is ρ-Lipschitz $(0 < \rho < \infty)$, and that $\|f\|_\infty \le C_b$ for any $f \in \mathcal{F}$ for some C_b. Let $C_\ell :=$*

$\sup_{t\in\{\pm 1\}}\ell(C_b^2, t)$. *For any $\delta > 0$, with probability at least $1 - \delta$,*

$$\mathrm{Er}_{\mathrm{clus}}(\mathrm{sign}\circ\widehat{f})$$

$$\leq \sqrt{\frac{1}{2}\psi^{-1}\left(\mathrm{Er}_{\mathrm{pair}}^\ell(f^*) + 4\rho\mathfrak{R}_{m_1}(\mathcal{F}) + \sqrt{\frac{2C_\ell^2\log\frac{2}{\delta}}{m_1}}\right)}.$$

Next, the class assignment error probability using pairwise supervision is analyzed.

Lemma 3. *Assume that $\pi_+ \neq \frac{1}{2}$. Let $\widehat{s}$ be the solution defined in Eq. (7). Then, we have*

$$\Pr\left(\widehat{s} \neq \underset{s\in\{\pm 1\}}{\arg\min} R_{\mathrm{point}}(s\cdot\mathrm{sign}\circ\widehat{f})\right)$$
$$\leq \exp\left(-\frac{m_2}{2}(2\pi_+ - 1)^2\big(2R_{\mathrm{point}}(\mathrm{sign}\circ\widehat{f}) - 1\big)^2\right).$$

Several observations from Lemma 3 follow. As $\pi_+ \to \frac{1}{2}$, the upper bound becomes looser. This comes from the fact that the estimation of the pointwise classification error with pairwise supervision becomes more difficult as $\pi_+ \to \frac{1}{2}$ (Shimada et al., 2021). Moreover, the discriminability of function $\widehat{f}$, i.e., $R_{\mathrm{point}}(\mathrm{sign}\circ\widehat{f})$, appears in the inequality and thus it is directly related to the error rate. Intuitively, if a given function classifies a large portion of data correctly, the optimal sign can be identified easily.

Finally, an overall excess risk bound is derived by combining Lemmas 2, 3, and the fact (10). Let $\mathrm{Er}_{\mathrm{point}}(h)$ denote the excess risk $R_{\mathrm{point}}(h) - R_{\mathrm{point}}^*$.

Theorem 3. *Suppose that we have $\pi_+ \neq \frac{1}{2}$. Let $r := \exp(-\frac{m_2}{2}(2\pi_+ - 1)^2(2R_{\mathrm{point}}(\mathrm{sign}\circ\widehat{f}) - 1)^2)$. Under the same assumptions as Lemma 2, for any $\delta > r$, with probability at least $1 - \delta$,*

$$\mathrm{Er}_{\mathrm{point}}(\widehat{s}\cdot\mathrm{sign}\circ\widehat{f})$$

$$\leq \sqrt{\frac{1}{2}\psi^{-1}\left(\mathrm{Er}_{\mathrm{pair}}^\ell(f^*) + 4\rho\mathfrak{R}_{m_1}(\mathcal{F}) + \sqrt{\frac{2C_\ell^2\log\frac{2}{\delta-r}}{m_1}}\right)}.$$

In the proof of Theorem 3, the surrogate excess risk $\mathrm{Er}_{\mathrm{pair}}^\ell(\widehat{f})$ is decomposed into the estimation error and the approximation error $\mathrm{Er}_{\mathrm{pair}}^\ell(f^*)$. If $\mathfrak{R}_{m_1}(\mathcal{F}) = o(1)$, the estimation error asymptotically vanishes and the upper bound approaches to the approximation error in probability. Under this condition, similarity learning successfully minimizes our desideratum $\mathrm{Er}_{\mathrm{point}}$, with a flexible enough $\mathcal{F}$ entailing the small approximation error. For example, linear-in-parameter model $\mathcal{F} = \{f(\boldsymbol{x}) = \boldsymbol{w}^\top\boldsymbol{\phi}(\boldsymbol{x}) + b\}$ satisfies $\mathfrak{R}_{m_1}(\mathcal{F}) = O(m_1^{-\frac{1}{2}})$ as shown in Kuroki et al. (2019, Lemma 5), where $\boldsymbol{w} \in \mathbb{R}^k$ and $b \in \mathbb{R}$ are weights and bias parameters

and $\boldsymbol{\phi}: \mathbb{R}^d \to \mathbb{R}^k$ are mapping functions. Note that our result is stronger than Zhang and Yan (2007) because they only provided the asymptotic convergence, while Theorem 3 provides a finite sample guarantee.

Discussion. Since class assignment admits the exponential decay of the error probability (Lemma 3) under the moderate condition ($\pi_+ \neq \frac{1}{2}$), we may set $m_2 \ll m_1$ in practice. In contrast, our excess risk bound of clustering error minimization (Lemma 2) is governed in part by ψ. The explicit rate depends on specific choices of loss functions: e.g., the hinge loss gives $\psi(u) = u$, and under the assumption $\mathfrak{R}_{m_1}(\mathcal{F}) = O(m_1^{-\frac{1}{2}})$, the explicit rate is $O_p(m_1^{-\frac{1}{4}})$.[8] This rate is no slower than the pointwisely supervised case $O_p(m^{-\frac{1}{2}})$ because $O(m^2)$ pairwise supervision can be generated with m pointwise labels.

Note again that CIPS assumes $\pi_+ \neq \frac{1}{2}$ only in class assignment (Step 2 & Lemma 3), not in clustering error minimization (Step 1 & Lemma 2). This is a subtle but notable difference from earlier similarity learning methods based on unbiased classification risk estimators, which requires $\pi_+ \neq \frac{1}{2}$ even in risk minimization (see Shimada et al. (2021)).

Our excess risk bound (Theorem 3) resembles transfer bounds among binary classification, class probability estimation (CPE), and bipartite ranking. Narasimhan and Agarwal (2013) reduced classification and CPE to ranking and showed that the excess risks of both classification and CPE can be upper-bounded by that of ranking. As can be seen in Narasimhan and Agarwal (2013), the excess risk of classification/CPE slows down to be $O(\lambda(m)^{-\frac{1}{2}})$ suppose that the excess risk of ranking is $\lambda(m)$. The same decay is observed in Theorem 3 as well, reducing classification to similarity learning. This decay $O((\cdot)^{-\frac{1}{2}})$ can be regarded as a *cost arising from problem reduction.*

5 Experiments

This section shows simulation results to confirm our findings: ⟨♣⟩ the sample complexity of the clustering error minimization via similarity learning (Lemma 2), ⟨♡⟩ the class-prior effect in similarity learning (Discussion in Section 4), and ⟨♠⟩ class assignment without pointwise supervision (Lemma 3). In addition, we compared with baselines using benchmark and real-world datasets (PubMed-Diabetes). All experiments except PubMed-Diabetes were carried out with 3.60GHz Intel® Core™ i7-7700 CPU and GeForce GTX 1070. Experiments

---

[8]As another example, the logistic loss gives $\psi(u) = \Omega(u^2)$, entailing the explicit rate $O_p(m_1^{-\frac{1}{8}})$ for the excess risk bound (Lemma 2). For more examples of ψ, see Steinwart (2007, Table 1).
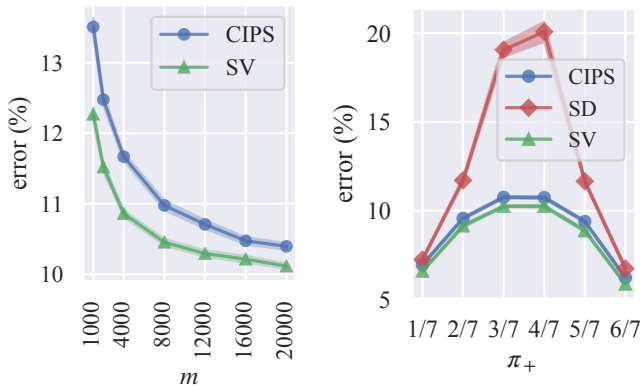
(a) ⟨♣⟩ $\pi_+ = 0.5$, various m (b) ⟨♡⟩ various π_+, $m = 10000$

Figure 1: (left) Mean clustering error and standard error (shaded areas) over 20 trials on MNIST. (right) Mean clustering error and standard error (shaded areas) over 10 trials on MNIST.
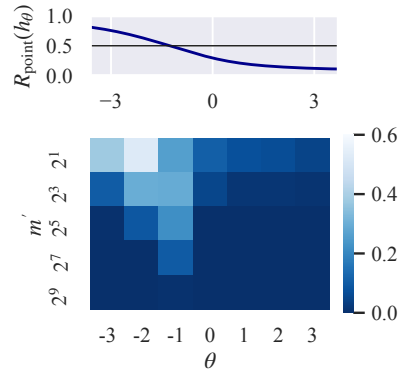

Figure 2: ⟨♠⟩ Classification error for each threshold classifier (upper) and the error probability of the proposed class assignment method over 10,000 trials (bottom) on the synthetic Gaussian dataset with $\pi_+ = 0.1$

Table 2: Mean clustering error and standard error on different benchmark datasets over 20 trials. Bold numbers indicate outperforming methods (excluding SV): among each configuration, the best one is chosen first, and then the comparable ones are chosen by one-sided t-test with the significance level 5%.

dataset (dim., π_+)	m	CIPS (Ours)	MCL	SD	OVPC	SSP	CKM	KM	(SV)
adult	100	39.8 (1.6)	38.4 (2.1)	30.8 (0.9)	45.0 (0.9)	**24.7 (0.3)**	28.9 (0.8)	**24.9 (0.5)**	21.9 (0.4)
(123, 0.24)	1000	**17.6 (0.3)**	**17.2 (0.3)**	20.5 (0.3)	45.5 (0.7)	24.2 (0.3)	27.9 (0.4)	27.9 (0.5)	15.9 (0.3)
codrna	100	**24.7 (1.8)**	32.3 (1.4)	**28.0 (1.3)**	32.0 (2.0)	45.5 (1.5)	46.7 (0.6)	42.5 (1.0)	11.0 (0.6)
(8, 0.33)	1000	**6.3 (0.2)**	**6.5 (0.2)**	8.8 (0.4)	28.3 (2.0)	44.8 (1.6)	46.1 (0.4)	45.4 (0.6)	6.3 (0.2)
ijcnn1	100	16.6 (2.3)	24.9 (2.9)	**10.7 (0.3)**	41.1 (1.1)	31.6 (2.0)	40.0 (1.3)	31.9 (2.4)	9.1 (0.2)
(22, 0.10)	1000	**7.7 (0.2)**	**7.9 (0.2)**	8.1 (0.2)	42.0 (1.4)	34.9 (1.7)	45.9 (0.8)	43.4 (0.7)	7.6 (0.2)
phishing	100	**12.7 (2.3)**	**12.8 (2.3)**	34.6 (1.8)	41.7 (1.0)	46.6 (0.5)	24.4 (3.4)	47.0 (0.5)	7.6 (0.2)
(44, 0.68)	1000	**6.5 (0.2)**	**6.3 (0.2)**	22.0 (1.0)	43.8 (1.1)	45.5 (0.5)	15.2 (2.7)	46.4 (0.5)	6.3 (0.2)
w8a	100	31.5 (1.9)	31.4 (2.1)	11.8 (0.3)	39.7 (1.4)	**5.3 (1.2)**	**6.8 (1.9)**	**5.5 (1.3)**	10.3 (0.4)
(300, 0.03)	1000	2.6 (0.2)	**2.2 (0.1)**	2.6 (0.2)	43.1 (0.8)	3.0 (0.1)	8.9 (2.6)	3.7 (0.5)	2.0 (0.1)

Table 3: Mean clustering error and standard error on Pubmed-Diabetes dataset over 20 trials. Bold numbers indicate outperforming method (excluding SV): chosen by the one-sided t-test in the same way as Table 2.

CIPS (Ours)	MCL	DML	(SV)
86.9 (0.4)	**86.6 (0.4)**	85.1 (0.2)	94.7 (0.1)

on PubMed-Diabetes were carried out with 1.40GHz Intel® Xeon Phi™ 7250. Full results are included in Appendix E. All simulation codes are available in the supplementary material.

Clustering error minimization on benchmark datasets. Tabular datasets from LIBSVM (Chang and Lin, 2011) and UCI (Dua and Graff, 2017) repositories and MNIST dataset (LeCun, 2013) were used in benchmarks. The labels of MNIST were binarized into even vs. odd digits. Pairwise supervision was generated by random coupling of pointwise data in the original datasets. We briefly introduce baselines below. Constrained k-means clustering (CKM) (Wagstaff et al., 2001) and semi-supervised spectral clustering (SSP) (Chen and Feng, 2012) are semi-supervised clustering methods based on k-means (MacQueen, 1967) and spectral clustering (von Luxburg, 2007), respectively. A method proposed by Zhang and Yan (2007) (OVPC) and similar-dissimilar classification (SD) (Shimada et al., 2021) are classification methods using pairwise supervision, which admit the generalization guarantee. Meta-classification likelihood (MCL) (Hsu et al., 2019) is an approach based on maximum likelihood estimation over pairwise labels. For reference, k-means clustering (KM) and supervised learning (SV) were compared. For classification methods that require model specification (i.e., CIPS, SD, MCL, OVPC, and SV), a linear model $f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} + b$ was used. For

CIPS, SD, and SV, we used the logistic loss, which is classification-calibrated. The rest of implementation details is deferred to Appendix E.

$\langle\clubsuit\rangle$ First, in order to verify the sample complexity behavior in Lemma 2, classifiers were trained with MNIST. The number of pairwise data m was set to each of $\{1,000, 2,000, 4,000, 8,000, 12,000, 16,000, 20,000\}$. Figure 1(a) presents the performances of CIPS and SV. This demonstrates that the clustering error of CIPS constantly decreases as m grows, which is consistent with Lemma 2. Moreover, CIPS performed more efficiently than expected in terms of sample complexity—as we discussed in Section 4, we expect that CIPS with $O(m^2)$ pairs performs comparably to SV with m data points.

$\langle\heartsuit\rangle$ Next, to see the effect of the class prior, we compared CIPS, SD, and SV with various class priors. In this experiment, train and test data were generated from MNIST under the controlled class prior π_+, where π_+ was set to each of $\{\frac{1}{7}, \ldots, \frac{6}{7}\}$. For each trial, 10,000 pairs were randomly subsampled from MNIST for training and the performance was evaluated with another 10,000 labeled examples. The average clustering errors and standard errors over ten trials are plotted in Figure 1(b). This result indicates that CIPS is less affected compared with SD.

Finally, we show the benchmark performances of each method on the tabular datasets in Table 2, where each cell contains the average clustering error and the standard error over 20 trials. For each trial, we subsampled $m \in \{100, 1000\}$ pairs for training data and 1,000 pointwise examples for evaluation. This result demonstrates CIPS performs better with large enough samples than most of the baselines and comparably to MCL. The performance difference between CIPS and clustering methods implies that larger samples do improve the downstream performance of CIPS thanks to its generalization guarantee (Theorem 3).

Class assignment on synthetic dataset. The performance of the proposed class assignment method was empirically investigated on synthetic dataset. The class-conditional distributions with the standard Gaussian distributions were used as the underlying distribution: $p(\boldsymbol{x}|y = +1) = \mathcal{N}(\boldsymbol{x}|\mu_+, \sigma_+)$ and $p(\boldsymbol{x}|y = -1) = \mathcal{N}(\boldsymbol{x}|\mu_-, \sigma_-)$. Throughout this experiment, we fixed $(\mu_+, \sigma_+, \mu_-, \sigma_-)$ to $(1, 1, -1, 2)$. Here, we consider a 1-D thresholded classifier denoted by $h_\theta(x) = 1$ if $x \geq \theta$ and -1 otherwise. Given the class prior $\pi_+ \in (0, 1)$, we generated m' pairwise examples from the above distributions and apply the proposed class assignment method for a fixed classifier h_θ. Then, we evaluated whether the estimated class assignment is optimal or not. Each parameter was set as follows: $m' \in \{2^1, 2^3, 2^5, 2^7, 2^9\}$, $\pi_+ = 0.1$, and $\theta \in \{-3, -2, \ldots, 3\}$.

For each (θ, π_+, m'), we repeated these data generation processes, class assignment, and evaluation procedure for 10,000 times.

$\langle\spadesuit\rangle$ The error probabilities are depicted in Figure 2. We find that the performance of the proposed class assignment method improves as (i) the number of pairwise examples m' grows and (ii) the classification error for a given classifier $R_{\mathrm{point}}(h_\theta)$ gets away from $\frac{1}{2}$. These results are aligned with our analysis in Section 4. Moreover, we observed that class assignment improves as the class prior π_+ becomes farther from $\frac{1}{2}$ in additional experiments in Appendix E.

Clustering error minimization on a real-world dataset. Finally, we show experimental results on a citation network dataset, PubMed-Diabetes.[9] The aim of this experiment is to verify that CIPS is robust enough against real-world noise in pairwise supervision.

We compare CIPS (proposed) with three baselines, MCL (described above), deep metric learning (DML), and SV (supervised). DML combines metric learning and k-means clustering: we first train embeddings so that their ℓ_2 distances are close for similar pairs and vice versa, and apply k-means clustering on the embeddings. More implementation details are deferred to Appendix E. The results are reported in Table 3, from which we can see that CIPS obtained a meaningful classifier even under the presence of real-world noise, and worked comparably to MCL and better than DML.

6 Conclusion

In this paper, we presented the underlying relationship between similarity learning and binary classification. Eventually, the two-step similarity learning procedure for binary classification with only pairwise supervision was obtained. Our similarity learning can elicit the underlying decision boundary and is less affected by the class prior. The post-processing class assignment is less costly than training a new classifier. Our framework can be related to many existing similarity learning methods with specific losses. It remains open to discuss the more flexible similarity model and the parallel connection for multi-class classification, in order to fully understand what knowledge we can elicit from similarity information.

---

[9]Available at https://linqs.soe.ucsc.edu/data.

References

F. R. Bach, G. R. Lanckriet, and M. I. Jordan. Multiple kernel learning, conic duality, and the smo algorithm. In *Proceedings of the 21st International Conference on Machine Learning*, page 6, 2004.

M.-F. Balcan, A. Blum, and N. Srebro. A theory of learning with similarity functions. *Machine Learning*, 72(1-2):89–112, 2008.

H. Bao, G. Niu, and M. Sugiyama. Classification from pairwise similarity and unlabeled data. In *Proceedings of the 35th International Conference on Machine Learning*, pages 461–470, 2018.

P. L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3 (Nov):463–482, 2002.

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

A. Bellet, A. Habrard, and M. Sebban. Similarity learning for provably accurate sparse linear classification. In *Proceedings of the 29th International Coference on Machine Learning*, pages 1491–1498, 2012.

M. Bilenko, S. Basu, and R. J. Mooney. Integrating constraints and metric learning in semi-supervised clustering. In *Proceedings of the 21st International Conference on Machine Learning*, pages 839–846, 2004.

J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah. Signature verification using a "siamese" time delay neural network. In *Advances in Neural Information Processing Systems 7*, pages 737–744, 1994.

C.-C. Chang and C.-J. Lin. LIBSVM: a library for support vector machines, 2011. URL http://www.csie.ntu.edu.tw/~cjlin/libsvm. ACM Transactions on Intelligent Systems and Technology.

W. Chen and G. Feng. Spectral clustering: A semi-supervised approach. *Neurocomputing*, 77:229–242, 2012.

Y. Chen, E. K. Garcia, M. R. Gupta, A. Rahimi, and L. Cazzanti. Similarity-based classification: Concepts and algorithms. *Journal of Machine Learning Research*, 10(3), 2009.

S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 539–546. IEEE, 2005.

C.-Y. Chuang, J. Robinson, Y.-C. Lin, A. Torralba, and S. Jegelka. Debiased contrastive learning. In *Advances in Neural Information Processing Systems 33*, pages 8765–8775, 2020.

T. Clanuwat, M. Bober-Irizar, A. Kitamoto, A. Lamb, K. Yamamoto, and D. Ha. Deep learning for classical Japanese literature. In *NeurIPS Workshop on Machine Learning for Creativity and Design*, 2018.

C. Cortes, M. Mohri, and A. Rostamizadeh. Two-stage learning kernel algorithms. In *Proceedings of the 27th International Conference on Machine Learning*, pages 239–246, 2010.

N. Cristianini, J. Shawe-Taylor, A. Elisseeff, and J. S. Kandola. On kernel-target alignment. In *Advances in Neural Information Processing Systems 15*, pages 367–373, 2002.

Z. Cui, N. Charoenphakdee, I. Sato, and M. Sugiyama. Classification from triplet comparison data. *Neural Computation*, 32(3):659–681, 2020.

S. Dan, H. Bao, and M. Sugiyama. Learning from noisy similar and dissimilar data. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 233–249, 2021.

J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon. Information-theoretic metric learning. In *Proceedings of the 24th International Conference on Machine Learning*, pages 209–216, 2007.

D. Dua and C. Graff. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

C. Dugas, Y. Bengio, F. Bélisle, C. Nadeau, and R. Garcia. Incorporating second-order functional knowledge for better option pricing. *Advances in Neural Information Processing Systems*, 13:472–478, 2000.

D. Eisenberg, E. M. Marcotte, I. Xenarios, and T. O. Yeates. Protein function in the post-genomic era. *Nature*, 405(6788):823–826, 2000.

A. Fahad, N. Alshatri, Z. Tari, A. Alamri, I. Khalil, A. Y. Zomaya, S. Foufou, and A. Bouras. A survey of clustering algorithms for big data: Taxonomy and empirical analysis. *IEEE Transactions on Emerging Topics in Computing*, 2(3):267–279, 2014.

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

R. Gomes, P. Welinder, A. Krause, and P. Perona. Crowdclustering. Advances in Neural Information Processing Systems 25, 2012.

W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, 2012.

Y.-C. Hsu, Z. Lv, J. Schlosser, P. Odom, and Z. Kira. Multi-class classification without multi-class labels. In *Proceedings of the 7th International Conference on Learning Representations*, 2019.

P. Kar and P. Jain. Similarity-based learning via data driven embeddings. In *Advances in Neural Information Processing Systems 24*, pages 1998–2006, 2011.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations*, 2015.

G. Koch, R. Zemel, and R. Salakhutdinov. Siamese neural networks for one-shot image recognition. In *ICML Deep Learning Workshop*, volume 2. Lille, 2015.

B. Kulis. Metric learning: A survey. *Foundations and Trends® in Machine Learning*, 5(4):287–364, 2013.

S. Kuroki, N. Charoenphakdee, H. Bao, J. Honda, I. Sato, and M. Sugiyama. Unsupervised domain adaptation based on source-guided discrepancy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4122–4129, 2019.

G. R. Lanckriet, N. Cristianini, P. Bartlett, L. E. Ghaoui, and M. I. Jordan. Learning the kernel matrix with semidefinite programming. *Journal of Machine Learning Research*, 5(Jan):27–72, 2004.

Y. LeCun. The MNIST database of handwritten digits, 2013. URL http://yann.lecun.com/exdb/mnist.

Z. Li and J. Liu. Constrained clustering by spectral kernel learning. In *IEEE 12th International Conference on Computer Vision*, pages 421–427, 2009.

L. Logeswaran and H. Lee. An efficient framework for learning sentence representations. In *Proceedings of the 6th International Conference on Learning Representations*, 2018.

J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. University of California Press, 1967.

T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*, pages 3111–3119, 2013.

M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of Machine Learning*. MIT Press, 2018.

H. Narasimhan and S. Agarwal. On the relationship between binary classification, bipartite ranking, and binary class probability estimation. In *Advances in Neural Information Processing Systems 26*, pages 2913–2921, 2013.

G. Niu, B. Dai, M. Yamada, and M. Sugiyama. Information-theoretic semi-supervised metric learning via entropy regularization. *Neural Computation*, 26(8):1717–1762, 2014.

K. Nozawa, P. Germain, and B. Guedj. PAC-Bayesian contrastive unsupervised representation learning. In *Conference on Uncertainty in Artificial Intelligence*, pages 21–30, 2020.

A. Okuno and H. Shimodaira. Hyperlink regression via Bregman divergence. *Neural Networks*, 126:362–383, 2020.

S. Ontañón. An overview of distance and similarity functions for structured data. *Artificial Intelligence Review*, 53(7):5309–5351, 2020.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

T. Sakai, M. C. du Plessis, G. Niu, and M. Sugiyama. Semi-supervised classification based on classification from positive and unlabeled data. In *Proceedings of the 34th International Conference on Machine Learning*, pages 2998–3006, 2017.

N. Saunshi, O. Plevrakis, S. Arora, M. Khodak, and H. Khandeparkar. A theoretical analysis of contrastive unsupervised representation learning. In *Proceedings of the 36th International Conference on Machine Learning*, pages 5628–5637, 2019.

T. Shimada, H. Bao, I. Sato, and M. Sugiyama. Classification from pairwise similarities/dissimilarities and unlabeled data via empirical risk minimization. *Neural Computation*, 33(5):1234–1268, 2021.

I. Steinwart. How to compare different loss functions and their risks. *Constructive Approximation*, 26(2):225–287, 2007.

J. Tang, M. Qu, M. Wang, M. Zhang, J. Yan, and Q. Mei. LINE: Large-scale information network embedding. In *Proceedings of the 24th International Conference on World Wide Web*, pages 1067–1077, 2015.

C. Tosh, A. Krishnamurthy, and D. Hsu. Contrastive estimation reveals topic posterior information to linear models. *Journal of Machine Learning Research*, 22(281):1–31, 2021.

N. Usunier, M. R. Amini, and P. Gallinari. Generalization error bounds for classifiers trained with interdependent data. In *Advances in Neural Information Processing Systems 18*, pages 1369–1376, 2005.

A. van den Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

B. van Rooyen, A. Menon, and R. C. Williamson. Learning with symmetric label noise: The importance of being unhinged. In *Advances in Neural Information Processing Systems 28*, pages 10–18, 2015.

R. Vogel, A. Bellet, and S. Clémençon. A probabilistic theory of supervised similarity learning for pointwise ROC curve optimization. In *Proceedings of the 35th International Conference on Machine Learning*, pages 5065–5074, 2018.

U. von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.

K. Wagstaff, C. Cardie, S. Rogers, and S. Schrödl. Constrained k-means clustering with background knowledge. In *Proceedings of the 18th International Conference on Machine Learning*, volume 1, pages 577–584, 2001.

L. Wang, M. Sugiyama, C. Yang, K. Hatano, and J. Feng. Theory and algorithm for learning with dissimilarity functions. *Neural Computation*, 21(5): 1459–1484, 2009.

X. Wang and A. Gupta. Unsupervised learning of visual representations using videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2794–2802, 2015.

K. Q. Weinberger and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10: 207–244, 2009.

S. Wu, X. Xia, T. Liu, B. Han, M. Gong, N. Wang, H. Liu, and G. Niu. Multi-class classification from noisy-similarity-labeled data. *arXiv preprint arXiv:2002.06508*, 2020.

H. Xiao, K. Rasul, and R. Vollgraf. Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

E. P. Xing, M. I. Jordan, S. J. Russell, and A. Y. Ng. Distance metric learning with application to clustering with side-information. In *Advances in Neural Information Processing Systems 16*, pages 521–528, 2003.

R. Yan, J. Zhang, J. Yang, and A. G. Hauptmann. A discriminative learning framework with pairwise constraints for video object classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4):578–593, 2006.

J. Zhang and R. Yan. On the value of pairwise constraints in classification and consistency. In *Proceedings of the 24th International Conference on Machine Learning*, pages 1111–1118. ACM, 2007.

A   Proofs of Theorems and Lemmas

In this section, we provide complete proofs for Theorem 1, Theorem 2, Lemma 1, Lemma 2, and Lemma 3.

A.1   Proof of Theorem 1

We derive an equivalent expression of the pairwise classification error R_{pair} as follows.

$$
\begin{aligned}
R_{\mathrm{pair}}(h) &= \underset{(X,Y)\sim p(\boldsymbol{x},y)}{\mathbb{E}}\ \underset{(X',Y')\sim p(\boldsymbol{x},y)}{\mathbb{E}} \left[\mathbb{1}\{h(X)\cdot h(X') \neq YY'\}\right] \\
&= \underset{(X,Y)\sim p(\boldsymbol{x},y)}{\mathbb{E}}\ \underset{(X',Y')\sim p(\boldsymbol{x},y)}{\mathbb{E}} \left[\mathbb{1}\{h(X)\neq Y\}\mathbb{1}\{h(X')=Y'\}\right] \\
&\quad + \underset{(X,Y)\sim p(\boldsymbol{x},y)}{\mathbb{E}}\ \underset{(X',Y')\sim p(\boldsymbol{x},y)}{\mathbb{E}} \left[\mathbb{1}\{h(X)=Y\}\mathbb{1}\{h(X)\neq Y'\}\right] \\
&= \underset{(X,Y)\sim p(\boldsymbol{x},y)}{\mathbb{E}} \left[\mathbb{1}\{h(X)\neq Y\}\right] \underset{(X',Y')\sim p(\boldsymbol{x},y)}{\mathbb{E}} \left[\mathbb{1}\{h(X')=Y'\}\right] \\
&\quad + \underset{(X,Y)\sim p(\boldsymbol{x},y)}{\mathbb{E}} \left[\mathbb{1}\{h(X)=Y\}\right] \underset{(X',Y')\sim p(\boldsymbol{x},y)}{\mathbb{E}} \left[\mathbb{1}\{h(X')\neq Y'\}\right] \\
&= 2 \underset{(X,Y)\sim p(\boldsymbol{x},y)}{\mathbb{E}} \left[\mathbb{1}\{h(X)\neq Y\}\right] \underset{(X',Y')\sim p(\boldsymbol{x},y)}{\mathbb{E}} \left[\mathbb{1}\{h(X')=Y'\}\right] \\
&= 2\,R_{\mathrm{point}}(h)\,(1-R_{\mathrm{point}}(h)).
\end{aligned}
\tag{11}
$$

We can transform the above equation as

$$
R_{\mathrm{point}}(h) = \frac{1}{2} \pm \frac{\sqrt{1-2R_{\mathrm{pair}}(h)}}{2}.
\tag{12}
$$

Then, we also have

$$
R_{\mathrm{point}}(-h) = 1 - R_{\mathrm{point}}(h) = \frac{1}{2} \mp \frac{\sqrt{1-2R_{\mathrm{pair}}(h)}}{2}.
\tag{13}
$$

By combining the results in Eqs. (12) and (13), we finally obtain Eq. (4), which completes the proof of Theorem 1. Remark that $0 \leq R_{\mathrm{pair}}(h) \leq \frac{1}{2}$ is evident from Eq. (11) because of $0 \leq R_{\mathrm{point}}(h) \leq 1$. $\qquad\square$

A.2   Proof of Theorem 2

The optimal sign s^* can be written as

$$
s^* = \underset{s\in\{\pm 1\}}{\arg\min}\, R_{\mathrm{point}}(s\cdot h) = \mathrm{sign}\left(R_{\mathrm{point}}(-h) - R_{\mathrm{point}}(h)\right).
\tag{14}
$$

According to Shimada et al. (2021), R_{point} is equivalently expressed as follows.

Lemma 4 (Theorem 1 in Shimada et al. (2021)). *Assume that $\pi_+ \neq \frac{1}{2}$. Then, the pointwise classification error for a given classifier $h : \mathcal{X} \rightarrow \mathcal{Y}$ can be equivalently represented as*

$$
\begin{aligned}
&R_{\mathrm{point}}(h) \\
&= \underset{(X,Y)\sim p(\boldsymbol{x},y)}{\mathbb{E}}\ \underset{(X',Y')\sim p(\boldsymbol{x},y)}{\mathbb{E}} \left[\frac{\mathbb{1}\{h(X)\neq YY'\} + \mathbb{1}\{h(X')\neq YY'\}}{2\,(2\pi_+ - 1)}\right] - \frac{1-\pi_+}{2\pi_+ - 1}.
\end{aligned}
\tag{15}
$$

By plugging Eq. (15) into Eq. (14), we obtain

$$
\begin{aligned}
&R_{\text{point}}(-h) - R_{\text{point}}(h) \\
&= \underset{(X,Y)\sim p(\boldsymbol{x},y)}{\mathbb{E}} \underset{(X',Y')\sim p(\boldsymbol{x},y)}{\mathbb{E}} \left[\frac{\mathbb{1}\left\{-h(X) \neq YY'\right\} + \mathbb{1}\left\{-h(X') \neq YY'\right\}}{2\left(2\pi_+ - 1\right)} \right] \\
&\quad - \underset{(X,Y)\sim p(\boldsymbol{x},y)}{\mathbb{E}} \underset{(X',Y')\sim p(\boldsymbol{x},y)}{\mathbb{E}} \left[\frac{\mathbb{1}\left\{h(X) \neq YY'\right\} + \mathbb{1}\left\{h(X') \neq YY'\right\}}{2\left(2\pi_+ - 1\right)} \right] \\
&= \underset{(X,Y)\sim p(\boldsymbol{x},y)}{\mathbb{E}} \underset{(X',Y')\sim p(\boldsymbol{x},y)}{\mathbb{E}} \left[\frac{1 - 2\cdot\mathbb{1}\left\{h(X) \neq YY'\right\} + 1 - 2\cdot\mathbb{1}\left\{h(X') \neq YY'\right\}}{2\left(2\pi_+ - 1\right)} \right] \\
&= \frac{1}{2\pi_+ - 1} \underset{(X,Y)\sim p(\boldsymbol{x},y)}{\mathbb{E}} \underset{(X',Y')\sim p(\boldsymbol{x},y)}{\mathbb{E}} \left[1 - \mathbb{1}\left\{h(X) \neq YY'\right\} - \mathbb{1}\left\{h(X') \neq YY'\right\} \right] \\
&= \frac{1}{2\pi_+ - 1}(1 - 2Q(h)).
\end{aligned}
\tag{16}
$$

Thus, we derive the following result.

$$
\begin{aligned}
s_h^* = \text{sign}\left(R_{\text{point}}(-h) - R_{\text{point}}(h)\right) &= \text{sign}\left(\frac{1}{2\pi_+ - 1}\right) \cdot \text{sign}(1 - 2Q(h)) \\
&= \text{sign}(2\pi_+ - 1) \cdot \text{sign}(1 - 2Q(h)),
\end{aligned}
\tag{17}
$$

which completes the proof of Theorem 2. Note that s^* can be either ± 1 when $Q(h) = \frac{1}{2}$, which is equivalent to $R_{\text{point}}(h) = R_{\text{point}}(-h) = \frac{1}{2}$. Here we arbitrarily set to $s^* = -\text{sign}(2\pi_+ - 1)$ in this case. $\qquad\square$

A.3 Proof of Lemma 1

We introduce the following notation:

$$
\begin{aligned}
S_{\text{point}}^\ell(\alpha, \eta) &:= \eta\ell(\alpha, +1) + (1 - \eta)\ell(\alpha, -1), \\
H_{\text{point}}^\ell(\eta) &:= \inf_{\alpha \in \mathbb{R}} S_{\text{point}}^\ell(\alpha, \eta), \\
H_{\text{point}}^{\ell,-}(\eta) &:= \inf_{\alpha:\alpha(2\eta-1)\leq 0} S_{\text{point}}^\ell(\alpha, \eta).
\end{aligned}
$$

S_{point}^ℓ represents the conditional ℓ-risk in the following sense:

$$
\underset{X}{\mathbb{E}}[S_{\text{point}}^\ell(f(X), p(Y = +1|X))] = R_{\text{point}}^\ell(f),
$$

where

$$
R_{\text{point}}^\ell(f) := \underset{(X,Y)\sim p(\boldsymbol{x},y)}{\mathbb{E}}[\ell(f(X), Y)].
$$

Define the function $\psi_{\text{point}} : [0, 1] \to [0, +\infty)$ by $\psi_{\text{point}} = \tilde{\psi}_{\text{point}}^{\star\star}$, where $\tilde{\psi}_{\text{point}}^{\star\star}$ is the Fenchel-Legendre biconjugate of $\tilde{\psi}_{\text{point}}$, and

$$
\tilde{\psi}_{\text{point}}(\varepsilon) := H_{\text{point}}^{\ell,-}\left(\frac{1+\varepsilon}{2}\right) - H_{\text{point}}^\ell\left(\frac{1+\varepsilon}{2}\right).
$$

ψ_{point} corresponds to ψ-transform introduced by Bartlett et al. (2006) exactly.

We will show that the statement of the lemma is satisfied by $\psi = \psi_{\text{point}}$ based on the *calibration analysis* (Steinwart,

2007). We further introduce the following notation:

$$
\begin{aligned}
S_{\mathrm{pair}}(\alpha, \alpha', \eta, \eta') := {} & \eta\eta' \mathbb{1}\{\operatorname{sign}(\alpha)\operatorname{sign}(\alpha') \neq +1\} \\
& + \eta(1 - \eta')\mathbb{1}\{\operatorname{sign}(\alpha)\operatorname{sign}(\alpha') \neq -1\} \\
& + (1 - \eta)\eta' \mathbb{1}\{\operatorname{sign}(\alpha)\operatorname{sign}(\alpha') \neq -1\} \\
& + (1 - \eta)(1 - \eta')\{\operatorname{sign}(\alpha)\operatorname{sign}(\alpha') \neq +1\}, \\
S_{\mathrm{pair}}^{\ell}(\alpha, \alpha', \eta, \eta') := {} & \eta\eta' \ell(\alpha\alpha', +1) + \eta(1 - \eta')\ell(\alpha\alpha, -1) \\
& + (1 - \eta)\eta'\ell(\alpha\alpha, -1) + (1 - \eta)(1 - \eta')\ell(\alpha\alpha, +1), \\
H_{\mathrm{pair}}(\eta, \eta') := {} & \inf_{\alpha, \alpha' \in \mathbb{R}} S_{\mathrm{pair}}(\alpha, \alpha', \eta, \eta'), \\
H_{\mathrm{pair}}^{\ell}(\eta, \eta') := {} & \inf_{\alpha, \alpha' \in \mathbb{R}} S_{\mathrm{pair}}^{\ell}(\alpha, \alpha', \eta, \eta').
\end{aligned}
$$

S_{pair}^{ℓ} represents the conditional ℓ-risk in the following sense:

$$
\mathop{\mathbb{E}}_{X,X'}[S_{\mathrm{pair}}^{\ell}(f(X), f(X'), p(Y = +1|X), p(Y' = +1|X'))] = R_{\mathrm{pair}}^{\ell}(f),
$$

and

$$
\mathop{\mathbb{E}}_{X,X'}[S_{\mathrm{pair}}(f(X), f(X'), p(Y = +1|X), p(Y' = +1|X'))] = R_{\mathrm{pair}}(\operatorname{sign}\circ f).
$$

Let $\tilde{\psi}_{\mathrm{pair}} : [0, 1] \to [0, +\infty)$ be the *calibration function* (Steinwart, 2007, Lemma 2.16) defined by

$$
\tilde{\psi}_{\mathrm{pair}}(\varepsilon) := \inf_{\eta,\eta\in[0,1]} \inf_{\alpha,\alpha'\in\mathbb{R}} S_{\mathrm{pair}}^{\ell}(\alpha, \alpha', \eta, \eta') - H_{\mathrm{pair}}^{\ell}(\eta, \eta')
$$

$$
\text{s.t. } S_{\mathrm{pair}}(\alpha, \alpha', \eta, \eta') - H_{\mathrm{pair}}(\eta, \eta') \geq \varepsilon.
$$

By the consequence of Lemma 2.9 of Steinwart (2007), $\tilde{\psi}_{\mathrm{pair}}(\varepsilon) > 0$ for all $\varepsilon > 0$ implies that $R_{\mathrm{pair}}^{\ell}(f) \to R_{\mathrm{pair}}^{*} \implies R_{\mathrm{pair}}(\operatorname{sign}\circ f) \to R_{\mathrm{pair}}^{*}$. Further, under this condition, Theorem 2.13 of Steinwart (2007) implies that $\tilde{\psi}_{\mathrm{pair}}$ is non-decreasing, invertible, and satisfies

$$
\tilde{\psi}_{\mathrm{pair}}^{\star\star}(R_{\mathrm{pair}}(\operatorname{sign}\circ f) - R_{\mathrm{pair}}^{*}) \leq R_{\mathrm{pair}}^{\ell}(f) - R_{\mathrm{pair}}^{\ell,*}
$$

for any measurable function f. Hence, it is sufficient to show that $\tilde{\psi}_{\mathrm{pair}}(\varepsilon) > 0$ for all $\varepsilon > 0$. Indeed, $\tilde{\psi}_{\mathrm{pair}} = \tilde{\psi}_{\mathrm{point}}$, and $\tilde{\psi}_{\mathrm{point}}(\varepsilon) > 0$ for all $\varepsilon > 0$ because ℓ is classification-calibrated (Bartlett et al., 2006, Lemma 2). From now on, we will see $\tilde{\psi}_{\mathrm{pair}} = \tilde{\psi}_{\mathrm{point}}$.

First, we simplify the constraint part of $\tilde{\psi}_{\mathrm{pair}}$. Since

$$
\begin{aligned}
S_{\mathrm{pair}}(\alpha, \alpha', \eta, \eta') = {} & (1 - \eta - \eta' + 2\eta\eta')\mathbb{1}\{\operatorname{sign}(\alpha)\operatorname{sign}(\alpha') = -1\} \\
& + (\eta + \eta' - 2\eta\eta')\mathbb{1}\{\operatorname{sign}(\alpha)\operatorname{sign}(\alpha') = +1\} \\
= {} & \tilde{\eta}\mathbb{1}\{\operatorname{sign}(\alpha)\operatorname{sign}(\alpha') = +1\} + (1 - \tilde{\eta})\mathbb{1}\{\operatorname{sign}(\alpha)\operatorname{sign}(\alpha') = -1\},
\end{aligned}
$$

where $\tilde{\eta} := 1 - \eta - \eta' + 2\eta\eta'$, we have $H_{\mathrm{pair}}(\eta, \eta') = \min\{\tilde{\eta}, 1 - \tilde{\eta}\}$. Similarly,

$$
S_{\mathrm{pair}}^{\ell}(\alpha, \alpha', \eta, \eta') = \tilde{\eta}\ell(\alpha\alpha', +1) + (1 - \tilde{\eta})\ell(\alpha\alpha', -1).
$$

With slight abuse of notation, we may write $S_{\mathrm{pair}}(\alpha, \alpha', \tilde{\eta}) = S_{\mathrm{pair}}(\alpha, \alpha', \eta, \eta')$ (same for S_{pair}^{ℓ}, H_{pair}, and H_{pair}^{ℓ}). By simple algebra, we obtain

$$
S_{\mathrm{pair}}(\alpha, \alpha', \tilde{\eta}) - H_{\mathrm{pair}}(\tilde{\eta}) = |2\tilde{\eta} - 1| \cdot \mathbb{1}\{(2\tilde{\eta} - 1)\operatorname{sign}(\alpha)\operatorname{sign}(\alpha') \leq 0\}.
$$

Noting that $\tilde{\eta}$ ranges over $[0, 1]$ with $\eta, \eta' \in [0, 1]$, we have

$$
\tilde{\psi}_{\mathrm{pair}}(\varepsilon) = \inf_{\tilde{\eta}\in[0,1]} \inf_{\alpha,\alpha'\in\mathbb{R}} S_{\mathrm{pair}}^{\ell}(\alpha, \alpha', \tilde{\eta}) - H_{\mathrm{pair}}^{\ell}(\tilde{\eta})
$$

$$
\text{s.t. } |2\tilde{\eta} - 1| \cdot \mathbb{1}\{(2\tilde{\eta} - 1)\operatorname{sign}(\alpha)\operatorname{sign}(\alpha') \leq 0\} \geq \varepsilon.
$$

If $\varepsilon = 0$, $\tilde{\psi}_{\text{pair}}(0) = 0$ and the infimum is attained by $\tilde{\eta} = \frac{1}{2}$ and arbitrary α and α'. If $\varepsilon > 0$, $\tilde{\eta} = \frac{1}{2}$ cannot satisfy the constraint. Hence, we assume $\tilde{\eta} \neq \frac{1}{2}$ from now on. When $\tilde{\eta} > \frac{1}{2}$, the constraint reduces to

$$\{\alpha\alpha' \leq 0 \wedge (\alpha, \alpha') \neq (0,0)\} \vee \tilde{\eta} \geq \frac{1+\varepsilon}{2}.$$

Since S^ℓ_{pair} contains α and α' only in the form of $\alpha\alpha'$, the infimum over $\{\alpha, \alpha' \in \mathbb{R} \mid \alpha\alpha' \leq 0 \wedge (\alpha, \alpha') \neq (0,0)\}$ is equal to that over $\{\alpha, \alpha' \in \mathbb{R} \mid \alpha\alpha' \leq 0\}$. If we write $\alpha\alpha' := \tilde{\alpha}$, then

$$\tilde{\psi}_{\text{pair}}(\varepsilon) = \inf_{\tilde{\eta} \in \left[\frac{1+\varepsilon}{2}, 1\right]} \inf_{\tilde{\alpha} \in \mathbb{R}: \tilde{\alpha} \leq 0} S^\ell_{\text{pair}}(\alpha, \alpha', \tilde{\eta}) - H^\ell_{\text{pair}}(\tilde{\eta})$$

$$= \inf_{\tilde{\eta} \in \left[\frac{1+\varepsilon}{2}, 1\right]} \inf_{\tilde{\alpha} \in \mathbb{R}: \tilde{\alpha} \leq 0} S^\ell_{\text{point}}(\tilde{\alpha}, \tilde{\eta}) - H^\ell_{\text{point}}(\tilde{\eta})$$

$$= \inf_{\tilde{\alpha} \in \mathbb{R}: \tilde{\alpha} \leq 0} S^\ell_{\text{point}}\left(\tilde{\alpha}, \frac{1+\varepsilon}{2}\right) - H^\ell_{\text{point}}\left(\frac{1+\varepsilon}{2}\right)$$

$$= H^{\ell,-}_{\text{point}}\left(\frac{1+\varepsilon}{2}\right) - H^\ell_{\text{point}}\left(\frac{1+\varepsilon}{2}\right)$$

$$= \tilde{\psi}_{\text{point}}(\varepsilon).$$

When $\tilde{\eta} < \frac{1}{2}$, $\tilde{\psi}_{\text{pair}} = \tilde{\psi}_{\text{point}}$ can be shown in the same way. Hence, the statement is proven. $\qquad\square$

A.4 Proof of Lemma 2

We start by introducing the following statement.

Lemma 5. *For real values α and β satisfying $0 \leq \alpha \leq \beta \leq 1$, we have*

$$\sqrt{\beta} - \sqrt{\alpha} \leq \sqrt{\beta - \alpha}. \tag{18}$$

Proof.

$$(\beta - \alpha) - (\sqrt{\beta} - \sqrt{\alpha})^2 = 2\sqrt{\alpha\beta} - 2\alpha = 2\sqrt{\alpha}\left(\sqrt{\beta} - \sqrt{\alpha}\right) \geq 0, \tag{19}$$

Thus we have $(\beta - \alpha) \geq (\sqrt{\beta} - \sqrt{\alpha})^2$, which completes the proof of Lemma 5. $\qquad\square$

With this lemma, an excess risk on clustering error can be connected with that on pairwise classification error as follows. From the equation in Eq. (4), we have

$$R^*_{\text{clus}} = \frac{1}{2} - \frac{\sqrt{1 - 2R^*_{\text{pair}}}}{2}. \tag{20}$$

Thus, we can bound excess risk on the clustering error as follows.

$$R_{\text{clus}}(\text{sign} \circ \widehat{f}) - R^*_{\text{clus}} = \left(\frac{1}{2} - \frac{\sqrt{1 - 2R_{\text{pair}}(\text{sign} \circ \widehat{f})}}{2}\right) - \left(\frac{1}{2} - \frac{\sqrt{1 - 2R^*_{\text{pair}}}}{2}\right)$$

$$= \frac{1}{2}\left\{\sqrt{1 - 2R^*_{\text{pair}}} - \sqrt{1 - 2R_{\text{pair}}(\text{sign} \circ \widehat{f})}\right\}$$

$$\leq \sqrt{\frac{R_{\text{pair}}(\text{sign} \circ \widehat{f}) - R^*_{\text{pair}}}{2}}$$

$$\leq \sqrt{\frac{1}{2}\psi^{-1}\left(R^\ell_{\text{pair}}(\widehat{f}) - R^{\ell,*}_{\text{pair}}\right)}, \tag{21}$$

where Lemma 5 and Lemma 1 were applied to obtain the penultimate and the last inequalities, respectively. The excess risk with respect to pairwise surrogate risk, i.e., $R^\ell_{\text{pair}}(\widehat{f}) - R^{\ell,*}_{\text{pair}}$, can be decomposed into *approximation error* and *estimation error* as

$$R^\ell_{\text{pair}}(\widehat{f}) - R^{\ell*}_{\text{pair}} = \underbrace{R^\ell_{\text{pair}}(f^*) - R^{\ell*}_{\text{pair}}}_{\text{approximation error}} + \underbrace{R^\ell_{\text{pair}}(\widehat{f}) - R^\ell_{\text{pair}}(f^*)}_{\text{estimation error}}, \tag{22}$$

where f^* is the minimizer of $R_{\text{pair}}^\ell(f)$ in a specified function space $\mathcal{F}$. Now, we provide the following upper bound for the estimation error with the Rademacher complexity.

Lemma 6. *Let $f^* \in \mathcal{F}$ be a minimizer of R_{pair}^ℓ, and $\widehat{f} \in \mathcal{F}$ be a minimizer of the empirical risk $\widehat{R}_{\text{pair}}^\ell$. Assume that the loss function ℓ is ρ-Lipschitz function with respect to the first argument $(0 < \rho < \infty)$, and all functions in the model class $\mathcal{F}$ are bounded, i.e., there exists an constant C_b such that $\|f\|_\infty \le C_b$ for any $f \in \mathcal{F}$. Let $C_\ell := \sup_{t \in \{\pm 1\}} \ell(C_b^2, t)$. For any $\delta > 0$, with probability at least $1 - \delta$,*

$$R_{\text{pair}}^\ell(\widehat{f}) - R_{\text{pair}}^\ell(f^*) \le 4\rho \mathfrak{R}_{m_1}(\mathcal{F}) + \sqrt{\frac{2C_\ell^2 \log \frac{2}{\delta}}{m_1}}. \tag{23}$$

Proof. The estimation error can be bounded as

$$\begin{aligned}
R_{\text{pair}}^\ell(\widehat{f}) - R_{\text{pair}}^\ell(f^*) &\le \left(R_{\text{pair}}^\ell(\widehat{f}) - \widehat{R}_{\text{pair}}^\ell(\widehat{f})\right) + \left(\widehat{R}_{\text{pair}}^\ell(f^*) - R_{\text{pair}}^\ell(f^*)\right) \\
&\le 2 \sup_{f \in \mathcal{F}} \left| R_{\text{pair}}^\ell(\widehat{f}) - \widehat{R}_{\text{pair}}^\ell(\widehat{f}) \right|.
\end{aligned} \tag{24}$$

With the Rademacher complexity, the following inequalities hold with probability at least $1 - \delta$.

$$\left| R_{\text{pair}}^\ell(\widehat{f}) - \widehat{R}_{\text{pair}}^\ell(\widehat{f}) \right| \le 2\mathfrak{R}_{m_1}(\ell \circ \mathcal{F}) + \sqrt{\frac{C_\ell^2 \log \frac{2}{\delta}}{2m_1}}, \tag{25}$$

where $\ell \circ \mathcal{F}$ indicates a class of composite functions defined by $\{\ell \circ f \mid f \in \mathcal{F}\}$. By applying Talagrand's lemma, the Rademacher complexity of $\ell \circ \mathcal{F}$ can be bounded as

$$\mathfrak{R}_{m_1}(\ell \circ \mathcal{F}) \le \rho \mathfrak{R}_{m_1}(\mathcal{F}). \tag{26}$$

The proofs of Eqs. (25) and (26) can be found in Mohri et al. (2018, Theorem 3.1 and Lemma 4.2), respectively. By plugging Eqs. (25) and (26) into Eq. (24), we obtain the result in Eq. (23). $\qquad\square$

By combining Eqs. (21), (22) and Lemma 6, we obtain the following inequality with probability at least $1 - \delta$,

$$R_{\text{clus}}(\text{sign} \circ \widehat{f}) - R_{\text{clus}}^* \le \sqrt{\frac{1}{2}\psi^{-1}\left(R_{\text{pair}}^\ell(f^*) - R_{\text{pair}}^{\ell,*} + 4\rho \mathfrak{R}_{m_1}(\mathcal{F}) + \sqrt{\frac{2C_\ell^2 \log \frac{2}{\delta}}{m_1}}\right)}. \tag{27}$$

$\qquad\square$

A.5 Proof of Lemma 3

We first derive a sufficient condition for the proposed class assignment fails. Let $\widehat{s}$ be a estimated class assignment for a given hypothesis $h : \mathcal{X} \to \mathcal{Y}$.

$$\begin{aligned}
\Pr\left(\widehat{s} \ne \operatorname*{arg\,min}_{s \in \{\pm 1\}} R_{\text{point}}(s \cdot h)\right) &= \Pr\left(\text{sign}\left(1 - 2\widehat{Q}(h)\right) \ne \text{sign}\left(1 - 2Q(h)\right)\right) \\
&= \begin{cases} \Pr\left(2\widehat{Q}(h) - 1 > 0\right) & (1 - 2Q(h) > 0), \\ \Pr\left(2\widehat{Q}(h) - 1 \le 0\right) & (\text{otherwise}) \end{cases} \\
&= \begin{cases} \Pr\left(\widehat{Q}(h) - Q(h) > \frac{1}{2} - Q(h)\right) & (1 - 2Q(h) > 0), \\ \Pr\left(Q(h) - \widehat{Q}(h) \ge Q(h) - \frac{1}{2}\right) & (\text{otherwise}) \end{cases}
\end{aligned} \tag{28}$$

By applying Hoeffding's inequality Hoeffding (1963), we obtain the following bounds.

$$\Pr\left(\widehat{Q}(h) - Q(h) > \frac{1}{2} - Q(h)\right) \le \exp\left(-2m_2\left(Q(h) - \frac{1}{2}\right)^2\right), \tag{29}$$

$$\Pr\left(Q(h) - \widehat{Q}(h) \ge Q(h) - \frac{1}{2}\right) \le \exp\left(-2m_2\left(Q(h) - \frac{1}{2}\right)^2\right), \tag{30}$$

where m_2 is the number of pairwise examples to compute $\widehat{Q}(h)$. Therefore, we can bound the error probability of the proposed class assignment method regardless of the value of $Q(h)$ as

$$\Pr\left(\widehat{s} \neq \underset{s \in \{\pm 1\}}{\arg\min} R_{\text{point}}(s \cdot h)\right) \leq \exp\left(-2m_2 \left(Q(h) - \frac{1}{2}\right)^2\right). \tag{31}$$

Now, we further explore how the term $Q(h) - \frac{1}{2}$ can be expressed. From the definition of Q and the equivalent risk expression in Eq. (15), we have

$$Q(h) = (2\pi_+ - 1)R_{\text{point}}(h) + 1 - \pi_+. \tag{32}$$

Therefore,

$$Q(h) - \frac{1}{2} = (2\pi_+ - 1)\left(R_{\text{point}}(h) - \frac{1}{2}\right). \tag{33}$$

By plugging Eq. (33) into Eq. (31), we finally obtain

$$\Pr\left(\widehat{s} \neq \underset{s \in \{\pm 1\}}{\arg\min} R_{\text{point}}(s \cdot h)\right) \leq \exp\left(-\frac{m_2}{2}(2\pi_+ - 1)^2 \left(2R_{\text{point}}(h) - 1\right)^2\right), \tag{34}$$

which completes the proof of Lemma 3. $\qquad\square$

B Discussion on Class Assignment

In this section, we discuss the impossibility of recovering the class assignment only with unlabeled validation data. Given a real-valued prediction function $f : \mathcal{X} \to \mathbb{R}$ and the class prior π_+, we consider the following class assignment strategy instead of the proposed method:

$$\tilde{s} := \text{sign}\,(2\pi_+ - 1) \cdot \text{sign}\left(\underset{X \sim p(\boldsymbol{x})}{\mathbb{E}}[\text{sign}\,(f(X))]\right). \tag{35}$$

Our aim is to estimate the optimal class assignment $s^* = \arg\min_{s \in \{\pm 1\}} R_{\text{point}}\,(s \cdot \text{sign} \circ f)$, which can be expressed by

$$
\begin{aligned}
s^* &= \text{sign}\,(R_{\text{point}}\,(-\,\text{sign} \circ f) - R_{\text{point}}\,(\text{sign} \circ f)) \\
&= \text{sign}\,((1 - R_{\text{point}}\,(\text{sign} \circ f)) - R_{\text{point}}\,(\text{sign} \circ f)) \\
&= \text{sign}\,(1 - 2R_{\text{point}}(\text{sign} \circ f)).
\end{aligned}
\tag{36}
$$

Thus, the following condition is necessary and sufficient for $\tilde{s} = s^*$:

$$\underbrace{\text{sign}\,(2\pi_+ - 1) \cdot \text{sign}\left(\underset{X \sim p(\boldsymbol{x})}{\mathbb{E}}[\text{sign}\,(f(X))]\right)}_{=\tilde{s}} \cdot \underbrace{\text{sign}\,(1 - 2R_{\text{point}}(\text{sign} \circ f))}_{=s^*} > 0. \tag{37}$$

We will investigate whether this condition always holds or not. Denote

$$R_{\text{point}}^+ := \pi_+ \underset{X \sim p(\boldsymbol{x}|y=+1)}{\mathbb{E}}[\mathbb{1}\,\{\text{sign}\,(f(X)) \neq +1\}], \tag{38}$$

$$R_{\text{point}}^- := (1 - \pi_+) \underset{X \sim p(\boldsymbol{x}|y=-1)}{\mathbb{E}}[\mathbb{1}\,\{\text{sign}\,(f(X)) \neq -1\}]. \tag{39}$$

Note that $0 \leq R_{\text{point}}^+ \leq \pi_+$ and $0 \leq R_{\text{point}}^- \leq 1 - \pi_+$ always hold. Now, we have

$$
\begin{aligned}
\mathop{\mathbb{E}}_{X \sim p(\boldsymbol{x})} \left[\mathbb{1}\left\{ \text{sign}\left(f(X)\right) = +1 \right\} \right] &= \pi_+ \mathop{\mathbb{E}}_{X \sim p(\boldsymbol{x}|y=+1)} \left[\mathbb{1}\left\{ \text{sign}\left(f(X)\right) = +1 \right\} \right] \\
&\quad + (1 - \pi_+) \mathop{\mathbb{E}}_{X \sim p(\boldsymbol{x}|y=-1)} \left[\mathbb{1}\left\{ \text{sign}\left(f(X)\right) = +1 \right\} \right] \\
&= \pi_+ \mathop{\mathbb{E}}_{X \sim p(\boldsymbol{x}|y=+1)} \left[\left(1 - \mathbb{1}\left\{ \text{sign}\left(f(X)\right) = -1 \right\} \right) \right] \\
&\quad + (1 - \pi_+) \mathop{\mathbb{E}}_{X \sim p(\boldsymbol{x}|y=-1)} \left[\mathbb{1}\left\{ \text{sign}\left(f(X)\right) = +1 \right\} \right] \\
&= -\pi_+ \mathop{\mathbb{E}}_{X \sim p(\boldsymbol{x}|y=+1)} \left[\mathbb{1}\left\{ \text{sign}\left(f(X)\right) = -1 \right\} \right] \\
&\quad + (1 - \pi_+) \mathop{\mathbb{E}}_{X \sim p(\boldsymbol{x}|y=-1)} \left[\mathbb{1}\left\{ \text{sign}\left(f(X)\right) = +1 \right\} \right] + \pi_+ \\
&= -R_{\text{point}}^+ + R_{\text{point}}^- + \pi_+.
\end{aligned}
\tag{40}
$$

Similarly, we have

$$
\mathop{\mathbb{E}}_{X \sim p(\boldsymbol{x})} \left[\mathbb{1}\left\{ \text{sign}\left(f(X)\right) = -1 \right\} \right] = R_{\text{point}}^+ - R_{\text{point}}^- + 1 - \pi_+.
\tag{41}
$$

By combining them, the following expression can be obtained.

$$
\begin{aligned}
\mathop{\mathbb{E}}_{X \sim p(\boldsymbol{x})} \left[\text{sign}\left(f(X)\right) \right] &= \mathop{\mathbb{E}}_{X \sim p(\boldsymbol{x})} \left[\mathbb{1}\left\{ \text{sign}\left(f(X)\right) = +1 \right\} \right] - \mathop{\mathbb{E}}_{X \sim p(\boldsymbol{x})} \left[\mathbb{1}\left\{ \text{sign}\left(f(X)\right) = -1 \right\} \right] \\
&= -2R_{\text{point}}^+ + 2R_{\text{point}}^- + 2\pi_+ - 1.
\end{aligned}
\tag{42}
$$

Hence, the necessary and sufficient condition (37) is rewritten as

$$
\text{sign}\left(2\pi_+ - 1\right) \cdot \text{sign}\left(-2R_{\text{point}}^+ + 2R_{\text{point}}^- + 2\pi_+ - 1\right) \cdot \text{sign}\left(1 - 2R_{\text{point}}^+ - 2R_{\text{point}}^-\right) > 0.
\tag{43}
$$

This condition is satisfied when π_+, R_{point}^+, and R_{point}^- satisfy any of the following conditions.

- $\pi_+ \geq \frac{1}{2}$, $R_{\text{point}}^- \geq R_{\text{point}}^+ + \frac{1}{2} - \pi_+$, and $R_{\text{point}}^- \leq -R_{\text{point}}^+ + \frac{1}{2}$,
- $\pi_+ \geq \frac{1}{2}$, $R_{\text{point}}^- < R_{\text{point}}^+ + \frac{1}{2} - \pi_+$, and $R_{\text{point}}^- > -R_{\text{point}}^+ + \frac{1}{2}$,
- $\pi_+ < \frac{1}{2}$, $R_{\text{point}}^- \geq R_{\text{point}}^+ + \frac{1}{2} - \pi_+$, and $R_{\text{point}}^- > -R_{\text{point}}^+ + \frac{1}{2}$,
- $\pi_+ < \frac{1}{2}$, $R_{\text{point}}^- < R_{\text{point}}^+ + \frac{1}{2} - \pi_+$, and $R_{\text{point}}^- \leq -R_{\text{point}}^+ + \frac{1}{2}$.
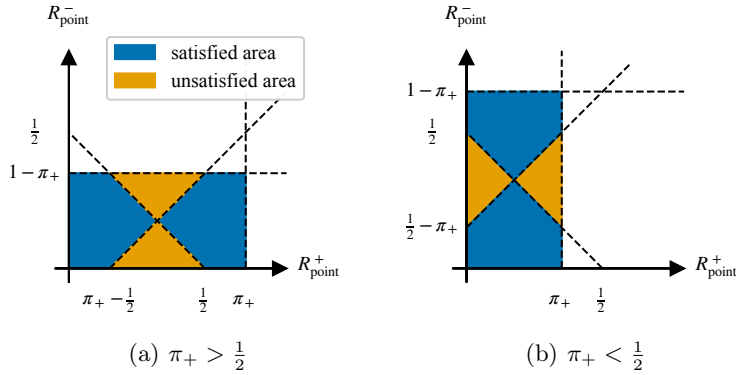

Figure 3: Illustration of the areas corresponding to the condition (43) (highlighted with blue). We have $\tilde{s} = s^*$ in the blue areas and otherwise in the orange areas.

The conditions (43) are depicted in Figure 3. As can be seen from this figure, for any binary classification problem (i.e., for any π_+), there exists a case where the class assignment with unlabeled data fails ($\tilde{s} \neq s^*$).

C Extension to semi-supervised learning

In real-world applications, we may face the situation where a large amount of unlabeled data are available along with pairwise data. Similarly to existing weakly-supervised classification frameworks such as positive-negative-unlabeled classification (Sakai et al., 2017) and similar-dissimilar-unlabeled classification (Shimada et al., 2021), we can easily incorporate unlabeled data for the estimation of R_{pair}^{ℓ}.

Theorem 4. *For non-negative real values* $(\gamma_1, \gamma_2, \gamma_3)$ *that satisfies* $\gamma_1 + \gamma_2 + \gamma_3 = 1$*, the risk* $R_{\text{pair}}^{\ell}(f)$ *can be equivalently expressed as:*

$$
\begin{aligned}
(\pi_+^2 + \pi_-^2) & \underset{(X,X') \sim p(\boldsymbol{x}, \boldsymbol{x}' | yy' = +1)}{\mathbb{E}} [(\gamma_1 + \gamma_2)\ell(f(X)f(X'), +1) - \gamma_2 \ell(f(X)f(X'), -1)] \\
+ 2\pi_+ \pi_- & \underset{(X,X') \sim p(\boldsymbol{x}, \boldsymbol{x}' | yy' = -1)}{\mathbb{E}} [(\gamma_1 + \gamma_3)\ell(f(X)f(X'), -1) - \gamma_3 \ell(f(X)f(X'), +1)] \\
+ & \underset{(X,X') \sim p(\boldsymbol{x}, \boldsymbol{x}')}{\mathbb{E}} [\gamma_3 \ell(f(X)f(X'), +1) + \gamma_2 \ell(f(X)f(X'), -1)],
\end{aligned}
\tag{44}
$$

where π_+ *and* π_- *denote positive and negative class proportions, respectively.*

With the expression in Eq. (44), we can use both pairwise supervision and unlabeled data for the empirical estimation of R_{pair}^{ℓ}. As well as the similar-unlabeled classification method (Bao et al., 2018), our method can be applied with only similar-unlabeled (or dissimilar-unlabeled) data by controlling parameters $(\gamma_1, \gamma_2, \gamma_3)$.

D Training with linear model and unhinged Loss

In general, the optimization problem in Eq. (6) is non-convex. Thus, it is not guaranteed whether we can achieve global optima with gradient descent. However, with specific model class and loss function, we can obtain an optimal solution more efficiently. Consider the linear model $f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x}$, where $\boldsymbol{w} \in \mathbb{R}^d$ are parameters. As a loss function, we consider the unhinged loss $\ell_{\text{UH}}(z, t) := 1 - tz$. This loss function is originally proposed in van Rooyen et al. (2015) to cope with label noises. Here we reformulate the optimization problem with linear model and the unhinged loss as follows.

$$
\widehat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w}} \widehat{R}_{\text{pair}}^{\ell_{\text{UH}}}(\boldsymbol{w}), \quad \text{s.t.} \quad \|\boldsymbol{w}\| = 1,
\tag{45}
$$

where

$$
\begin{aligned}
\widehat{R}_{\text{pair}}^{\ell_{\text{UH}}}(\boldsymbol{w}) &:= \frac{1}{m} \sum_{(X,X',T) \in \mathfrak{D}_1} \left(1 - T\boldsymbol{w}^\top X \cdot \boldsymbol{w}^\top X'\right) \\
&= 1 - \boldsymbol{w}^\top \left(\frac{1}{m} \sum_{(X,X',T) \in \mathfrak{D}_1} TXX'^\top\right) \boldsymbol{w} \\
&= 1 - \boldsymbol{w}^\top \left(\frac{1}{2m} \sum_{(X,X',T) \in \mathfrak{D}_1} T\left(XX'^\top + X'X^\top\right)\right) \boldsymbol{w} \\
&= 1 - \boldsymbol{w}^\top M \boldsymbol{w},
\end{aligned}
\tag{46}
$$

where M denotes the Hermitian matrix $\frac{1}{2m} \sum_{(X,X',T) \in \mathfrak{D}_1} T\left(XX'^\top + X'X^\top\right)$. The constraint $\|\boldsymbol{w}\| = 1$ is necessary to prevent the objective function from divergence. Let $\lambda_1, \ldots, \lambda_d$ be eigenvalues of the matrix M that satisfies $\lambda_1 \geq \cdots \geq \lambda_d$, and $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_d$ be corresponding eigenvectors that satisfy $\|\boldsymbol{v}_i\| = 1$ for all $i \in \{1, \ldots, d\}$. The following statement is known as a property of *Rayleigh quotient* (Horn and Johnson, 2012).

$$
\boldsymbol{v}_1 = \arg\max_{\boldsymbol{w} \in \mathbb{R}^d, \|\boldsymbol{w}\| = 1} \boldsymbol{w}^\top M \boldsymbol{w}.
\tag{47}
$$

Thus, the analytical solution of the constrained optimization problem in Eq. (45) is obtained as

$$
\widehat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w} \in \mathbb{R}^d, \|\boldsymbol{w}\| = 1} \widehat{R}_{\text{pair}}^{\ell_{\text{UH}}}(\boldsymbol{w}) = \arg\max_{\boldsymbol{w} \in \mathbb{R}^d, \|\boldsymbol{w}\| = 1} \boldsymbol{w}^\top M \boldsymbol{w} = \boldsymbol{v}_1.
\tag{48}
$$

E   Full version of experimental results

In this section, we show the implementation details and the full versions of experimental results in Section 5, which were omitted in the main body due to the limited space.

implementation details (clustering error minimization on benchmark datasets).   The implementation details of our method (CIPS) and each baseline were as follows.

- CIPS (Ours): The empirical pairwise classification risk R_{pair} (6) was computed with the logistic loss. The linear model $f(\boldsymbol{x}) = \boldsymbol{w}^{\top}\boldsymbol{x} + b$ was used. The risk was optimized with the stochastic gradient descent (minibatch size: 64 / learning rate: 10^{-2} / ℓ_2-regularization parameter: 10^{-4} / training epochs: 500).

- MCL (Hsu et al., 2019): The loss function is based on the maximum likelihood, that is, the logistic loss as in the original paper. The model and optimization setup were the same as CIPS.

- SD (Shimada et al., 2021): Their proposed classification risk was computed with the logistic loss. The model and optimization setup were the same as CIPS.

- OVPC (Zhang and Yan, 2007): We followed the authors to use the squared loss and the closed-form minimizer was evaluated.

- SSP (von Luxburg, 2007): Pairwise data were used as hard constraints. In order to construct the neighborhood sets for the Laplacian matrix, 5-nearest neighbors were used. The features are obtained by constraints propagation. In order to perform the final k-means clustering on the obtained features, scikit-learn implementation (Pedregosa et al., 2011) was used with the default parameters.

- CKM (Wagstaff et al., 2001): Pairwise data were used as hard constraints. Clustering was carried out with 10 different random initializations and the best one was reported. For each initialization, the number of maximum iterations was set to 300 and the tolerance parameter was set to 10^{-4}.

- KM (MacQueen, 1967): Pairwise data were used for training without all link information. Scikit-learn implementation (Pedregosa et al., 2011) of k-means clustering was used with the default parameters.

- SV (Supervised): The true class labels were revealed during training. The model and optimization setup were the same as CIPS.

Implementation details (clustering error minimization on a real-world dataset).   Pubmed-Diabetes dataset is a citation network dataset consists of 19,717 nodes representing scientific publications related to diabetes and 44,338 (directed) edges representing citing relationships. Each node is described by 500-dimensional TF/IDF features, and categorized into three classes, among which we pick class 1 ("Diabetes Mellitus, Experimental") and 3 ("Diabetes Mellitus Type 2") to convert it into a binary-labeled dataset.

The implementation details of our method and the baselines were as follows.

- CIPS (Ours): The 4-layer perceptron (500-8-8-8-1) with the softplus activation (Dugas et al., 2000) was used. The softmax cross entropy was optimized with Adam (Kingma and Ba, 2015) (minibatch size: 4,096 / learning rate: 10^{-3} / training epochs: 100). The ℓ_2-regularization parameter is chosen from $\{10^{-2}, 10^{-4}, 10^{-6}\}$ by the five-fold cross-validation. The early stopping is applied with the patience of 10 epochs. We randomly extracted 20% of the nodes as test data. The pairwise supervision was generated as follows: first extracted the edges whose both ends are in the training data as similar, then randomly coupled the non-connected nodes as dissimilar, with the same numbers of similar and dissimilar pairs. About 19,000 pairs were obtained.

- MCL (Hsu et al., 2019): The setup of model, optimization, and data generation was the same as CIPS.

- DML (Chopra et al., 2005): The metric loss function proposed by Chopra et al. (2005) was used. The model was the same as CIPS except the last layer, and 8-dimensional outputs of the penultimate layer were used as the embeddings, on which k-means clustering was performed. Scikit-learn implementation (Pedregosa et al., 2011) of k-means clustering was used with the default parameters. The setup of optimization and data generation was the same as CIPS.

- SV (Supervised): Labeled 7,889 nodes ($\pi_+ \approx 0.65$) were used during training. The setup of model and optimization was the same as CIPS.

Full results. Table 4 shows the performance comparison with baseline methods on ten datasets from UCI and LIBSVM repositories. Figure 4 presents the sample complexity of our method on three image classification datasets including MNIST (LeCun, 2013), Fashion-MNIST (Xiao et al., 2017), and Kuzushiji-MNIST (Clanuwat et al., 2018), where the original ten class categories were converted into positive/negative labels by grouping even/odd class labels. Figure 5 demonstrates the performance of our class assignment method with various class priors $\pi_+ \in \{0.1, 0.4, 0.7\}$.

Table 4: Mean clustering error and standard error on different benchmark datasets over 20 trials. Bold numbers indicate outperforming methods, chosen by one-sided t-test with the significance level 5%.

| dataset (dim., π_+)	m	CIPS (Ours)	MCL	SD	OVPC	SSP	CKM	KM	(SV)
adult	100	39.8 (1.6)	38.4 (2.1)	30.8 (0.9)	45.0 (0.9)	**24.7 (0.3)**	28.9 (0.8)	**24.9 (0.5)**	21.9 (0.4)
(123, 0.24)	500	21.5 (1.0)	**19.3 (0.4)**	23.2 (0.4)	44.7 (0.9)	24.3 (0.3)	28.2 (0.4)	27.5 (0.5)	16.9 (0.3)
	1000	**17.6 (0.3)**	**17.2 (0.3)**	20.5 (0.3)	45.5 (0.7)	24.2 (0.3)	27.9 (0.4)	27.9 (0.5)	15.9 (0.3)
banana	100	**43.6 (0.6)**	**44.5 (0.6)**	45.3 (0.6)	46.0 (0.7)	**43.0 (1.0)**	46.4 (0.7)	45.8 (0.7)	44.6 (0.6)
(2, 0.45)	500	43.1 (0.8)	43.3 (0.6)	45.1 (0.7)	46.0 (0.7)	**14.3 (0.7)**	45.5 (0.6)	44.4 (0.4)	45.1 (0.6)
	1000	44.4 (0.6)	44.3 (0.7)	44.4 (0.5)	46.2 (0.5)	**11.0 (0.2)**	45.0 (0.7)	44.0 (0.3)	45.1 (0.7)
codrna	100	**24.7 (1.8)**	32.3 (1.4)	**28.0 (1.3)**	32.0 (2.0)	45.5 (1.5)	46.7 (0.6)	42.5 (1.0)	11.0 (0.6)
(8, 0.33)	500	**6.4 (0.2)**	10.6 (0.3)	12.0 (0.6)	28.0 (2.1)	48.6 (0.3)	46.2 (0.3)	44.0 (0.7)	6.6 (0.2)
	1000	**6.3 (0.2)**	**6.5 (0.2)**	8.8 (0.4)	28.3 (2.0)	44.8 (1.6)	46.1 (0.4)	45.4 (0.6)	6.3 (0.2)
ijcnn1	100	16.6 (2.3)	24.9 (2.9)	**10.7 (0.3)**	41.1 (1.1)	31.6 (2.0)	40.0 (1.3)	31.9 (2.4)	9.1 (0.2)
(22, 0.10)	500	**7.7 (0.2)**	8.2 (0.2)	8.3 (0.2)	41.6 (1.3)	33.0 (2.5)	45.4 (0.8)	41.7 (0.7)	7.9 (0.2)
	1000	**7.7 (0.2)**	7.9 (0.2)	**8.1 (0.2)**	42.0 (1.4)	34.9 (1.7)	45.9 (0.8)	43.4 (0.7)	7.6 (0.2)
magic	100	**24.9 (1.3)**	**28.7 (1.8)**	30.7 (1.3)	41.9 (1.0)	47.1 (0.5)	45.5 (1.2)	44.0 (1.2)	21.8 (0.4)
(10, 0.35)	500	**21.5 (0.3)**	**21.3 (0.3)**	25.5 (0.8)	39.6 (1.5)	46.8 (0.5)	46.8 (0.4)	44.4 (0.4)	20.8 (0.3)
	1000	**21.3 (0.3)**	**20.9 (0.3)**	23.8 (0.4)	39.5 (1.7)	43.6 (0.9)	46.8 (0.3)	44.6 (0.4)	20.7 (0.3)
phishing	100	**12.7 (2.3)**	**12.8 (2.3)**	34.6 (1.8)	41.7 (1.0)	46.6 (0.5)	24.4 (3.4)	47.0 (0.5)	7.6 (0.2)
(44, 0.68)	500	7.2 (0.2)	**6.6 (0.1)**	26.9 (1.4)	42.9 (0.8)	46.0 (0.5)	16.9 (2.6)	46.4 (0.5)	6.5 (0.2)
	1000	**6.5 (0.2)**	**6.3 (0.2)**	22.0 (1.0)	43.8 (1.1)	45.5 (0.5)	15.2 (2.7)	46.4 (0.5)	6.3 (0.2)
phoneme	100	**28.2 (1.2)**	33.1 (1.9)	**29.1 (1.2)**	38.4 (1.3)	31.0 (1.3)	**28.0 (1.0)**	32.9 (1.2)	25.7 (0.4)
(5, 0.71)	500	**25.0 (0.4)**	**24.2 (0.5)**	26.1 (0.6)	38.6 (1.9)	25.5 (0.5)	28.0 (0.8)	32.7 (0.3)	25.0 (0.3)
	1000	**25.2 (0.4)**	**25.0 (0.4)**	26.0 (0.4)	39.8 (1.5)	**24.5 (0.5)**	30.2 (0.6)	32.7 (0.3)	25.3 (0.2)
spambase	100	**13.8 (1.0)**	**13.3 (1.3)**	31.6 (1.5)	39.7 (1.3)	40.5 (0.4)	**15.9 (2.0)**	39.7 (1.3)	10.5 (0.3)
(57, 0.39)	500	9.4 (0.2)	**8.6 (0.2)**	22.6 (0.9)	38.0 (1.6)	40.8 (0.3)	11.5 (0.2)	37.4 (2.3)	8.5 (0.2)
	1000	8.3 (0.2)	**7.6 (0.1)**	19.7 (0.8)	39.3 (1.2)	40.2 (0.4)	11.5 (0.2)	39.7 (1.3)	7.8 (0.2)
w8a	100	31.5 (1.9)	31.4 (2.1)	11.8 (0.3)	39.7 (1.4)	**5.3 (1.2)**	**6.8 (1.9)**	**5.5 (1.3)**	10.3 (0.4)
(300, 0.03)	500	5.6 (0.7)	4.2 (0.5)	**3.2 (0.1)**	38.3 (1.3)	**3.5 (0.1)**	14.0 (3.1)	5.5 (1.1)	2.6 (0.1)
	1000	2.6 (0.2)	**2.2 (0.1)**	2.6 (0.2)	43.1 (0.8)	3.0 (0.1)	8.9 (2.6)	3.7 (0.5)	2.0 (0.1)
waveform	100	**18.2 (0.3)**	**17.7 (0.3)**	26.4 (0.9)	41.9 (1.6)	44.1 (0.6)	41.0 (1.3)	45.1 (0.6)	16.2 (0.2)
(21, 0.33)	500	15.8 (0.2)	**15.1 (0.2)**	20.2 (0.5)	38.9 (1.3)	44.9 (0.7)	45.1 (0.6)	47.1 (0.4)	14.8 (0.2)
	1000	**14.9 (0.2)**	**14.7 (0.2)**	18.4 (0.3)	37.0 (1.7)	45.5 (0.5)	44.9 (0.5)	47.8 (0.4)	14.4 (0.2)
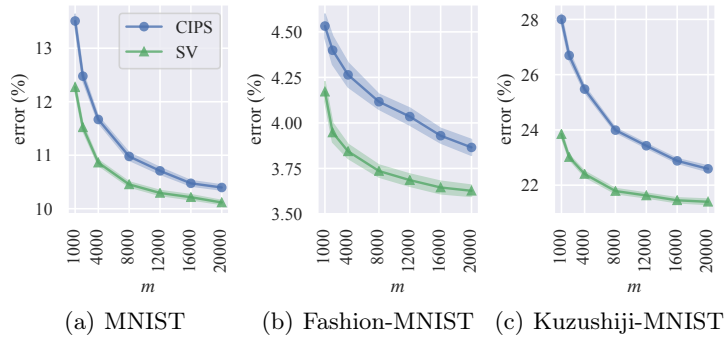
Figure 4: Mean clustering error and standard error (shaded areas) over 20 trials on image classification datasets.
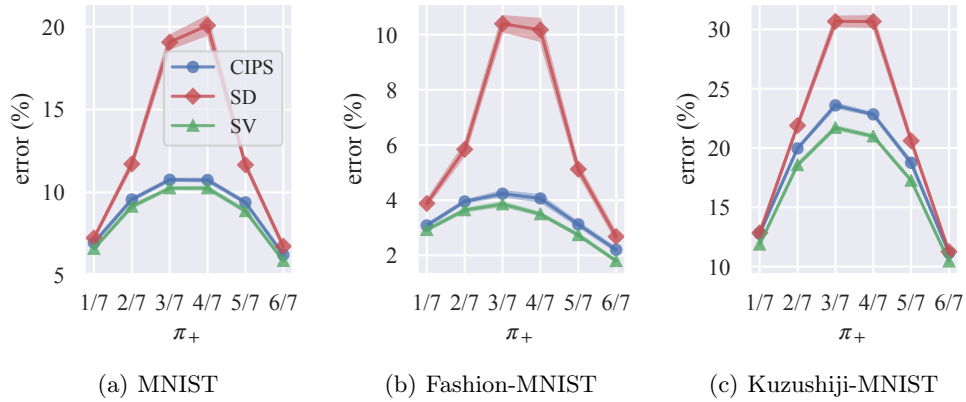

Figure 5: Mean clustering error and standard error (shaded areas) over ten trials on image classification datasets under controlled class priors.
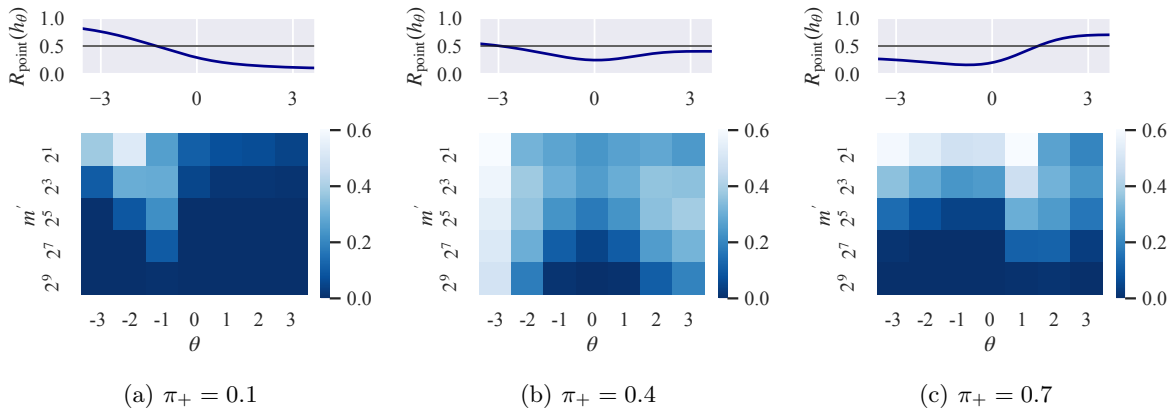

Figure 6: Classification error for each threshold classifier (upper) and the error probability of the proposed class assignment method over 10,000 trials (bottom) on the synthetic Gaussian dataset with $\pi_+ \in \{0.1, 0.4, 0.7\}$. The detail of the dataset is described in Section 5.