Labels, Information, and Computation: Efficient Learning Using Sufficient Labels

Shiyu Duan   MICHAELSHIYU3@GMAIL.COM
Spencer Chang   CHANG.SPENCER@UFL.EDU
José C. Príncipe   PRINCIPE@CNEL.UFL.EDU
Department of Electrical and Computer Engineering
University of Florida
Gainesville, FL 32611, USA

Editor: Isabelle Guyon

Abstract

In supervised learning, obtaining a large set of fully-labeled training data is expensive. We show that we do not always need full label information on every single training example to train a competent classifier. Specifically, inspired by the principle of sufficiency in statistics, we present a statistic (a summary) of the fully-labeled training set that captures almost all the relevant information for classification but at the same time is easier to obtain directly. We call this statistic "sufficiently-labeled data" and prove its sufficiency and efficiency for finding the optimal hidden representations, on which competent classifier heads can be trained using as few as a single randomly-chosen fully-labeled example per class. Sufficiently-labeled data can be obtained from annotators directly without collecting the fully-labeled data first. And we prove that it is easier to directly obtain sufficiently-labeled data than obtaining fully-labeled data. Furthermore, sufficiently-labeled data is naturally more secure since it stores relative, instead of absolute, information. Extensive experimental results are provided to support our theory.

Keywords: Classification, Deep Learning, Data Efficiency, Data Security, Privacy-Preserving Learning

1. Introduction

Consider a c-class classification task with i.i.d. training data $\{X_i, Y_i\}_{i=1}^n$, with each X_i being an input example and $Y_i \in \{1, \ldots, c\}$ its label. Do we always need full label information on every single training example to estimate a competent classifier on its distribution? Despite that such practice is pervasive in machine learning, we think that more efficient strategies exist.

Our inspiration originates from the principle of sufficiency in statistics. On a high level, a sufficient statistic is a function of data that contains all its information when it comes to estimating an unknown parameter of the underlying distribution (Casella and Berger, 2002). Effective data reduction can be achieved with a sufficient statistic at hand. For example, for normal distribution, sample mean, a one-number summary, is just as informative as the whole sample for estimating population mean. And similarly, sample variance suffices for estimating population variance. For uniform distributions, sample minimum and maximum

together carries an equal amount of information as the whole sample for estimating the support.

In this work, we extend this principle of sufficiency to classification with a focus on reducing the cost of obtaining labeled training data. We discover a statistic on fully-labeled data that (1) can be obtained directly from annotators without having to collect fully-labeled data first and (2) is easier to collect compared to fully-labeled data but at the same time (3) captures relevant information from data for learning the optimal hidden representations. Then a classifier head can be trained on the learned hidden representations using a set of fully-labeled data. Competent classifier heads can be trained using as few as a single randomly-chosen fully-labeled example per class. This allows one to learn performant models with less costly training data by using mostly this statistic in place of fully-labeled data. We call this statistic "sufficiently-labeled data" to emphasize the connection. And the sufficiency of this statistic can be rigorously proven in certain settings, allowing us to interpret training with sufficiently-labeled data as training using only relevant information, no more and no less.

If given a fully-labeled dataset, encoding it into a sufficiently-labeled one can be viewed as a generic encryption scheme that protects data security and user privacy. Specifically, a sufficiently-labeled training set contains only "relative labels" on user pairs. And no absolute information on individuals can be re-identified based on these relative labels. This is ideal in settings where the labels to be predicted on individuals contain sensitive information such as the diagnosis of a certain disease. In this case, sensitive user information no longer needs to be stored and/or transported, enhancing the security of the pipeline without extra overhead or compromise in performance.

Contributions

This work makes the following theoretical contributions. We define sufficiently-labeled data to be input example pairs with each pair having a binary label stating whether the two examples are from the same class. We then present a framework to learn with a mixture of sufficiently-labeled and fully-labeled data. Specifically, the hidden layers are trained using sufficiently-labeled data, and the linear output layer is then trained using fully-labeled data without fine-tuning the hidden layers.[1] We prove the following results.

1. Sufficiently-labeled data is sufficient for finding the optimal parameters of the hidden layers, and the proposed learning method produces solutions that are as competent as those produced using fully-labeled data.

2. Having more sufficiently-labeled data reduces the need for fully-labeled data.

3. Sufficiently-labeled data can be derived from fully-labeled data, but it can also be directly collected. And it is easier to collect sufficiently-labeled data than fully-labeled data.

In addition to our theoretical contributions, we present the following experimental results. On MNIST, Fashion-MNIST (Xiao et al., 2017), SVHN (Netzer et al., 2011), CIFAR-10, and CIFAR-100 (Krizhevsky and Hinton, 2009), we empirically verify that, given a large

---

1. Many popular classifier networks such as the ResNets (He et al., 2016) have a linear output layer, so our method is widely applicable.

set of sufficiently-labeled data for training the network body, only a single randomly-selected fully-labeled example per class is needed to train a state-of-the-art classifier. Further, our proposed learning method shares similar sample complexity as learning using purely fully-labeled data.

Paper Structure

In Section 2 and 3, we present our theory of sufficient labels. How sufficient labels can be used to help protect user privacy is described in Section 4. Related work is then discussed in Section 5. Section 6 contains our experimental results. And finally, conclusions are given in Section 7.

2. A Theoretical Framework for Learning From Sufficient Labels

In this chapter, we present a framework for learning with a mixture of sufficiently-labeled and fully-labeled data. The settings and notations are established first in Section 2.1. A learning method is presented in Section 2.2. We give an overview on the theoretical guarantees we provide for this method in Section 2.3. Then we proceed to formally present and prove these results in Section 2.4, 2.5, and 2.6.

2.1 Notations and Settings

In this section, some notations are established and the settings in which our theoretical results work are described. We also formally define our sufficient labels. Note that the settings below are only for the theoretical results in this section. We later generalize our learning algorithm to the general classification setup. In particular, we first present and analyze our learning algorithm in the binary classification setting. The algorithm is later generalized to multi-class classification and we prove its efficacy through extensive experiments, but rigorously extending the theory to the multi-class case is beyond the scope of this paper. With that said, a theoretical analysis on the multi-class case should still be strongly related to our treatment on the binary case in this paper. This is because the multi-class extension of our algorithm essentially treats the multi-class classification problem as a set of binary classification subproblems using the classic one-versus-the-rest strategy, and solves it using a group of one-versus-the-rest subclassifiers (this aligns with the standard practice in designing deep neural network classifiers, where each output layer node is responsible for predicting one particular class).

2.1.1 Some Notations

We use $\mathbb{1}$ to denote the indicator function. The support of a function or random element f is denoted $\mathrm{supp}(f)$. The cardinality of a set $\mathbb{U}$ is written as $|\mathbb{U}|$. For data, we use upper case letters for random elements and lower case letters for realizations of random elements. Throughout the text, the term "with probability" will be abbreviated to "w.p.".

2.1.2 Data

We consider two-class classification with data being random elements $X \in \mathbb{X} \subseteq \mathbb{R}^d, Y \in \{+, -\}$, where X is the input and Y its label, and $\mathbb{X}$ is a subspace. Let the density function

of (X, Y) be $p_{X,Y}(x, y)$. For any two independent $(X_1, Y_1), (X_2, Y_2)$, it is evident that

$$p_{X_1, X_2 | Y_1, Y_2}(x_1, x_2) = p_{X_1 | Y}(x_1) p_{X_2 | Y}(x_2). \tag{1}$$

Let $\alpha = \Pr(Y = +) \in (0, 1)$.

2.1.3 SUFFICIENT LABELS

Definition 1 (Sufficient Label) *Given random variables Y, Y', a sufficient label is defined as*

$$T(Y, Y') = \mathbb{1}(Y = Y'). \tag{2}$$

Definition 2 (Sufficiently-Labeled Data) *A sufficiently-labeled dataset of size $n \in \mathbb{N}$ is given as*

$$\{(X_{1_i}, X_{2_i}, T(Y_{1_i}, Y_{2_i}))\}_{i=1}^{n}, \tag{3}$$

where $\{(X_{j_i}, Y_{j_i})\}_{j=1,2,i=1,\dots,n}$ is a set of i.i.d. random elements sharing distribution with (X, Y).

Sufficiently-labeled data can be derived from fully-labeled data. On a fully-labeled sample of size two $(X_1, Y_1), (X_2, Y_2)$, a sufficient label is a summary of the labels. Specifically, it summarizes whether these two examples are from the same class. In general, if we have a fully-labeled sample $\{(X_i, Y_i)\}_{i=1}^{2n}$, we can reduce it into a sufficiently-labeled sample that is a summary of the original sample:

$$\{(X_{1_i}, X_{2_i}, T(Y_{1_i}, Y_{2_i}))\}_{i=1}^{n}. \tag{4}$$

The reduction can be performed in multiple ways by arranging the indices $1_i, 2_i$.

On the other hand, in practice, a sufficiently-labeled sample can evidently be obtained directly without having to collect the fully-labeled dataset first. For this, the annotators should be presented with example pairs and should be asked to label if each pair are from the same class. The actual class of each individual example in any pair does not have to be identified. Conceptually, this requires less effort from annotators than asking them to fully label each individual example. Moreover, as a theoretical result, we later justify from a learning theoretical point of view why it is easier to collect sufficiently-labeled data.

2.1.4 MODEL AND RISK

We consider models of the form $f = f_2 \circ F_1$, where $\mathbb{F}_1 \ni F_1 : \mathbb{X} \to \mathbb{H}$, $\mathbb{F}_2 \ni f_2 : \mathbb{H} \to \mathbb{R}$. Our theory works with any real inner product space $\mathbb{H}$. For typical deep learning settings, we may consider $\mathbb{H}$ to be $\mathbb{R}^p$ for some p for convenience with the dot product and the ℓ^2 norm as the inner product and the norm of choice. $\mathbb{F}_1, \mathbb{F}_2$ are some hypothesis spaces. For binary classification, the model prediction on X is obtained as $\text{sign}(f(X))$. Define

$$\mathbb{F} = \{f | f = f_2 \circ F_1, f_2 \in \mathbb{F}_2, F_1 \in \mathbb{F}_1\}. \tag{5}$$

In addition, assume that f_2 is a linear model in $\mathbb{R}^p$: $f_2(\cdot) = \langle w, \phi(\cdot) \rangle$, where ϕ is some feature map, and that it is parameterized by w. ϕ and w are assumed to satisfy

$$\|\phi(u)\| = r, \forall u, \|w\| \le \frac{1}{r}. \tag{6}$$

Note that this model formulation is broad enough to include the popular classifier networks such as the VGG networks (Simonyan and Zisserman, 2014), ResNets (He et al., 2016), and DenseNets (Huang et al., 2017). For these models, ϕ can be the nonlinearity between the network body and the final linear layer. If needed, one may normalize the network activation vector after this nonlinearity and the weight vector of the output layer to ensure that Eq. 6 is satisfied. For example, suppose we are given activation vector ϕ', weight vector w', then we may set $\phi = \phi'r/\|\phi'\|$ and $w = w'/(\|w'\|r)$.

The regular hinge loss is given as:

$$\ell^{0+} : \mathbb{R} \times \{+, -\} \to \mathbb{R} : (\hat{y}, y) \mapsto \max(0, 1 - y\hat{y}). \tag{7}$$

Based on the assumption on f (Eq. 6), we have

$$|\hat{y}| = |f(x)| = |\langle w, \phi(F_1(x))\rangle| \le \|w\|\|\phi(F_1(x))\| \le 1. \tag{8}$$

Therefore, it is equivalent to analyze this unbounded (from below) loss:

$$\ell : \mathbb{R} \times \{+, -\} \to \mathbb{R} : (\hat{y}, y) \mapsto -y\hat{y}. \tag{9}$$

Define the risk to be

$$R(f_2 \circ F_1, X, Y) = E_{X,Y}\ell(f_2 \circ F_1(X), Y). \tag{10}$$

The goal is to minimize this risk. A sample mean estimation to this risk is given as follows.

$$\hat{R}(f_2 \circ F_1, \{(x_i, y_i)\}_{i=1}^n) = \frac{1}{n}\sum_{i=1}^n \ell(f_2 \circ F_1(x_i), y_i), \tag{11}$$

where $\{(x_i, y_i)\}_{i=1}^n$ is a realization of an i.i.d. random sample sharing the same distribution as (X, Y).

2.2 How to Learn With Sufficient Labels

We present an algorithm for learning with sufficiently-labeled data. A similar learning method was proposed for modularizing deep learning workflows (Duan et al., 2020), but was not analyzed from the perspective of sufficient labels.

Suppose we are given the hypothesis space $\mathbb{F}$. Also, suppose we are given training data $\left\{\left(x_{1_i}, x'_{1_i}, T\left(y_{1_i}, y'_{1_i}\right)\right)\right\}_{i=1}^{n_1}$, $\{x_{2_i}, y_{2_i}\}_{i=1}^{n_2}$, with $\left(x_{i_j}, y_{i_j}\right)$ and $\left(x'_{1_j}, y'_{1_j}\right)$ being i.i.d. sharing the same distribution as (X, Y) for all i, j.

Define

$$\ell_1\left(F_1(x), F_1\left(x'\right), T\left(y, y'\right)\right) = (-1)^{T(y,y')+1}\left\|\phi \circ F_1(x) - \phi \circ F_1\left(x'\right)\right\|^2. \tag{12}$$

And also define a sample mean estimation

$$\hat{R}_1\left(F_1, \left\{\left(x_{1_i}, x'_{1_i}, T\left(y_{1_i}, y'_{1_i}\right)\right)\right\}_{i=1}^{n_1}\right) = \frac{1}{n_1}\sum_{i=1}^{n_1} \ell_1\left(F_1\left(x_{1_i}\right), F_1\left(x'_{1_i}\right), T\left(y_{1_i}, y'_{1_i}\right)\right), \tag{13}$$

The learning algorithm consists of the following two steps.

i Step 1 (Training the Hidden Layers): Find an $\hat{F}_1$ in

$$\underset{F_1 \in \mathbb{F}_1}{\arg\min} \hat{R}_1 \left(F_1, \left\{ \left(x_{1_i}, x'_{1_i}, T\left(y_{1_i}, y'_{1_i} \right) \right) \right\}_{i=1}^{n_1} \right). \tag{14}$$

ii Step 2 (Training the Output Layer): Find an $\hat{f}_2$ in

$$\underset{f_2 \in \mathbb{F}_2}{\arg\min} \hat{R} \left(f_2 \circ \hat{F}_1, \{(x_{2_i}, y_{2_i})\}_{i=1}^{n_2} \right). \tag{15}$$

iii Returns: $\hat{f}_2 \circ \hat{F}_1$.

Note that in Step 2, F_1 is kept frozen at $\hat{F}_1$. Also note that Step 1 only requires sufficiently-labeled data and Step 2 fully-labeled data.

Some immediate questions follow. (1) How do we know if and how well this learning algorithm works in terms of finding

$$\underset{f_2 \circ F_1 \in \mathbb{F}}{\arg\min} R(f_2 \circ F_1, X, Y)? \tag{16}$$

(2) Is it easier to collect training data for this algorithm than learning with only fully-labeled data? These questions will be rigorously answered in the following sections. Before diving into the detailed arguments, we first formally present the claims we wish to prove for this algorithm in the next section. This serves as an overview on our theoretical results that helps facilitate a better high-level understanding.

2.3 Overview of Theoretical Results

Suppose empirical risk minimization (ERM) on $\hat{R}$ (learning with purely fully-labeled data) can map training data $\{(x_i, y_i)\}_{i=1}^{n}$ and a hypothesis space $\mathbb{F}$ to a solution $F \in \mathbb{F}$ that attains at most γ true risk. We can make the following claims for our learning algorithm $\mathcal{A}_{\text{ours}}$ in Section 2.2.

1. (Section 2.4) $\mathcal{A}_{\text{ours}}$ can map $\left\{ \left(x_{1_i}, x'_{1_i}, T\left(y_{1_i}, y'_{1_i} \right) \right) \right\}_{i=1}^{n_1}$, $\{(x_{2_i}, y_{2_i})\}_{i=1}^{n_2}$, and $\mathbb{F}$ to a hypothesis $F_{ours} \in \mathbb{F}$ with at most γ true risk for some n_1, n_2.

2. (Section 2.5) n_2 decreases with n_1. In other words, having more sufficiently-labeled data reduces the need for fully-labeled data. This result is important because it shows that we can not only use sufficient labels *with* full labels, but also use sufficient labels *in place of* full labels.

3. (Section 2.6) Given X, X', sufficient label $T(Y, Y')$ is easier to obtain than full labels Y, Y'. This can be justified from several perspectives. For example, sufficient label is conceptually easier to annotate and may simply be more suitable in certain scenarios. From a learning theoretical standpoint, the sample complexity of learning a classifier to predict full label exceeds that for sufficient label as the number of underlying classes increase.

For the first claim, we empirically show that $n_1 + n_2$ is $\mathcal{O}(n)$. This shows that $\mathcal{A}_{\text{ours}}$ is as performant in finding the risk minimizer. The third claim is a general one on the sufficient labels, meaning that it does not depend on the proposed learning algorithm or the settings assumed in this section. These three statements will be proven in the subsequent three sections, respectively.

On a high level, these results together yields the following conclusions. First, performant models can be trained efficiently with our learning algorithm using a mixture of sufficiently-labeled and fully-labeled training data. In addition, it is easier to collect labeled training data for our learning method compared to learning with only fully-labeled data.

2.4 Finding the Risk Minimizer With Sufficient Labels

To justify that the proposed learning algorithm can indeed find a minimizer for the true risk R given enough training data, we can show that its Step 1 finds an $F_1 \in \mathbb{F}_1$ for which there exists an $f_2 \in \mathbb{F}_2$ such that $f_2 \circ F_1$ is a risk minimizer. To this end, we have the following theorem. The proofs of all our theoretical results are provided in the supplementary materials.

Theorem 3 *Define*

$$\mathbb{D} = \underset{F_1 \in \mathbb{F}_1}{\arg\min} \, E_{X,X'|Y \neq Y'} \left(- \left\| \phi \circ F_1(X) - \phi \circ F_1(X') \right\|^2 \right), \tag{17}$$

$$\mathbb{S} = \{ F_1 \in \mathbb{F}_1 | \Pr(\phi \circ F_1(X) = \phi \circ F_1(X')|Y = Y') \tag{18}$$
$$= \Pr(\phi \circ F_1(X) \neq \phi \circ F_1(X')|Y \neq Y') = 1\},$$

$$\mathbb{F}^\star = \left\{ f_2 \circ F_1 \middle| f_2 \circ F_1 \in \underset{f_2 \in \mathbb{F}_2, F_1 \in \mathbb{F}_1}{\arg\min} \, R(f_2 \circ F_1, X, Y) \right\}, \tag{19}$$

$$\mathbb{F}_1^\star = \{ F_1 | \exists f_2, f_2 \circ F_1 \in \mathbb{F}^\star \}. \tag{20}$$

If $\mathbb{D} \cap \mathbb{S} \neq \emptyset$ and $|\text{supp}(\phi \circ F_1(X))| > 4, \forall F_1 \notin \mathbb{S}$, then

$$\mathbb{F}_1^\star = \mathbb{D} \cap \mathbb{S}. \tag{21}$$

Now, to show that the proposed learning method finds a true risk minimizer, it suffices to show that, given enough data, Step 1 finds an element in $\mathbb{D} \cap \mathbb{S}$. To see this, note that

$$\frac{1}{n_1} \sum_{i=1}^{n_1} \mathbb{1} \left(T \left(y_{1_i}, y'_{1_i} \right) = 0 \right) \left(- \left\| \phi \circ F_1 \left(x_{1_i} \right) - \phi \circ F_1 \left(x'_{1_i} \right) \right\|^2 \right) \tag{22}$$

can be seen as an approximation to

$$E_{X,X'|Y \neq Y'} \left(- \left\| \phi \circ F_1(X) - \phi \circ F_1 \left(X' \right) \right\|^2 \right), \tag{23}$$

and

$$\frac{1}{n_1} \sum_{i=1}^{n_1} \mathbb{1} \left(T \left(y_{1_i}, y'_{1_i} \right) = 1 \right) \left\| \phi \circ F_1 \left(x_{1_i} \right) - \phi \circ F_1 \left(x'_{1_i} \right) \right\|^2 \tag{24}$$

to

$$E_{X,X'|Y = Y'} \left\| \phi \circ F_1(X) - \phi \circ F_1 \left(X' \right) \right\|^2, \tag{25}$$

7

which is minimized if and only if $\phi \circ F_1(X) = \phi \circ F_1(X')$ w.p. 1 given $Y = Y'$. More formally, since the data pairs are assumed to be i.i.d., the law of large numbers directly applies here and states that the empirical risks (Eq. 22 and Eq. 24) eventually converge to their expectation counterparts (Eq. 23 and Eq. 25) as n_1 grows. More analysis on finite sample behavior can be done here, and other forms of empirical risks that enjoy better convergence properties may exist. Further study on these topics is out of the scope of this paper and we leave it as future work.

2.4.1 CONNECTING SUFFICIENTLY-LABELED DATA WITH SUFFICIENT STATISTIC

In Theorem 3, $\mathbb{F}_1^\star$ is reformulated with expressions involving only sufficiently-labeled data $(X, X', T(Y, Y'))$. This implies that $(X, X', T(Y, Y'))$ contains as much information as (X, Y) for finding $\mathbb{F}_1^\star$, which is analogous to how a sufficient statistic supplies all the information for estimating an unknown parameter (Casella and Berger, 2002).

This similarity between sufficiently-labeled data and a sufficient statistic is summarized as follows.

• **For estimating an unknown parameter of a statistical model, a sufficient statistic is as informative as the entire sample (Casella and Berger, 2002).**

• **For finding the optimal classifier body, sufficiently-labeled data is as informative as fully-labeled data.**

2.4.2 INTERPRETATIONS: RETHINKING CLASSIFICATION IN TERMS OF PAIRWISE STATISTIC

In classification, data examples are usually considered individually and used with full label information. However, our results state that pairwise statistics of data contain all relevant information for learning the optimal hidden representations, allowing for data reduction. Intuitively, this is easy to understand. Consider F_1 to be learning a pattern in a feature space such that f_2, a linear classifier in that feature space, can most effectively classify (in terms of minimizing the hinge loss). Then as illustrated in Fig. 1, how F_1 arranges each individual example does not matter — all patterns are equally optimal if (1) each pair from distinct classes are mapped as far away from each other as possible and (2) each pair from the same class are mapped to the same point. This pattern is fully described using pairwise summary and that an optimal F_1 is one that learns such a pattern is essentially what Theorem 3 states. Indeed, the set $\mathbb{S}$ characterizes the set of F_1 that, with probability 1, maps every example from the same class to the same vector representation and those from distinct classes to different ones. On the other hand, $\mathbb{D}$ requires that F_1 maximizes (in expectation) the distance between example pairs from different classes.

This sufficiency of $(X, X', T(Y, Y'))$ allows us to understand it as a way to partition the sample space and leave only relevant information (Casella and Berger, 2002). Thus, learning with sufficiently-labeled examples directly is learning using relevant data, no more and no less. And going from fully-labeled data to sufficiently-labeled can be interpreted as extracting relevant information prior to any training.

Finally, from the perspective of learning representations, $\phi \circ F_1$ determines the learned internal representations of a network. Hence, our results show that internal representations
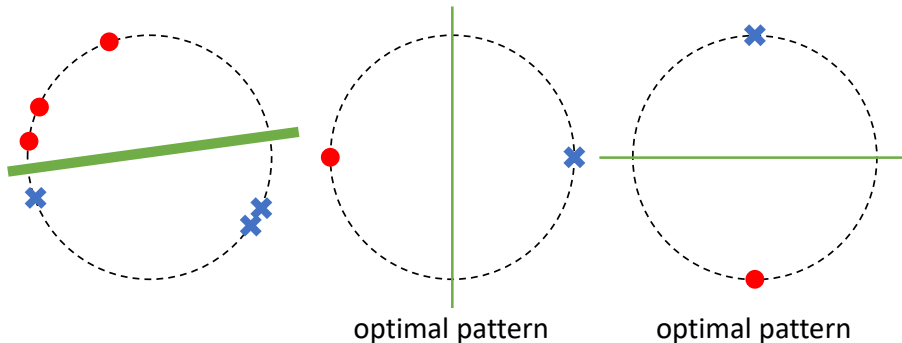
Figure 1: The sufficiency of sufficiently-labeled data. The hidden module F_1 learns to generate a pattern in a feature space (illustrated here) such that the output layer f_2, a linear model in this feature space, can most effectively classify. Due to the assumption on the feature map, i.e., $\|\phi(u)\| = r, \forall u$, every example must be on a circle in this illustration. For each pattern, the f_2 that achieves zero hinge loss (perfect separation) with minimum $\|w\|$ (model capacity) is the green line. The weight of the line is proportional to $\|w\|$. The optimal pattern, i.e., one that allows perfect separation with the smallest possible capacity (and therefore best generalization (Vapnik, 2013; Shalev-Shwartz and Ben-David, 2014)), is the one where each pair of examples from the same class are mapped to the same point, whereas each pair from different classes are as far away as possible. This optimality is described only using pairwise relationships and sufficient labels.

for classification can be learned using only sufficiently-labeled data. Fully labeled data is only needed to find the optimal linear mapping into the label space.

2.4.3 SAMPLE COMPLEXITY

Theorem 3 guarantees that the proposed learning algorithm finds a risk minimizer given enough sufficiently-labeled and fully-labeled training data, but it does not say anything about the sample complexity of the algorithm. We will later show via experiments that this algorithm in fact has similar sample complexity as one that leverages purely fully-labeled data. In other words, if ERM can use n fully-labeled examples to find a solution with a certain test performance, the proposed algorithm can find a solution with similar test performance using a mixture of sufficiently-labeled and fully-labeled data, with total sample size being comparable to n.

2.5 Having More Sufficiently-Labeled Data Means Needing Fewer Fully-Labeled Data

The proposed learning algorithm requires two sets of training data, one sufficiently-labeled, the other fully-labeled, in this order. In this section, we show that, to attain a particular

test performance, the needed number of fully-labeled training examples decreases with the number of sufficiently-labeled examples used. To this end, we have the following theorem.

Theorem 4 *Define $\mathbb{F}_{2,A} = \{f_2 | \|w\| \leq A\}$, and order these hypothesis spaces with A. Let $\mathbb{S}, \mathbb{D}$ be defined as in Theorem 3 with the assumptions therein satisfied. And let a true risk value $\gamma < 0$ be given.[2] For each $F_1 \in \mathbb{F}_1$ such that there exists $f_2 \in \mathbb{F}_2$ with $f_2 \circ F_1$ attaining this true risk value, denote the smallest $\mathbb{F}_{2,A}$ such that $\min_{f_2 \in \mathbb{F}_{2,A}} R(f_2 \circ F_1, X, Y) = \gamma$ as $\mathbb{F}_{2,F_1,\gamma}$. For each such F_1, let an $n_2 \in \mathbb{N}$ be given and define*

$$\hat{\mathbb{F}}^{\star}_{2,F_1,\gamma} = \operatorname*{arg\,min}_{f_2 \in \mathbb{F}_{2,F_1,\gamma}} \hat{R}\left(f_2 \circ F_1, \{(x_{2_i}, y_{2_i})\}_{i=1}^{n_2}\right). \tag{26}$$

Let $t(F_1)$ be the function that gives the smallest A such that for all $f_2 \in \mathbb{F}_{2,F_1,\gamma}$, we have $\|w\| \leq A$. For any given probability δ, we have, with probability at least $1 - \delta$, that

$$\sup_{\hat{f}_2 \in \hat{\mathbb{F}}^{\star}_{2,F_1,\gamma}} R\left(\hat{f}_2 \circ F_1, X, Y\right) - \gamma \leq 2\frac{t(F_1)\,r}{\sqrt{n_2}} + 5t(F_1)\,r\sqrt{\frac{2\ln(8/\delta)}{n_2}}. \tag{27}$$

And

$$\operatorname*{arg\,min}_{F_1 \in \mathbb{F}_1} t(F_1) = \mathbb{S} \cap \mathbb{D}. \tag{28}$$

To understand this theorem, first note that the norm upper bound A returned by $t(F_1)$ cannot be arbitrarily small. To see this, note that for an F_1, $\mathbb{F}_{2,F_1,\gamma}$ is defined to contain an f_2 such that $f_2 \circ F_1$ attains true risk value γ. This implicitly requires that the $\|w\|$ of this f_2 cannot be arbitrarily small. Indeed, note that the true risk of this $f_2 \circ F_1$ is $-E_{X,Y} Y f_2 \circ F_1(X)$ and this term can be lower bounded by $\|w\|$ multiplied by some constant using Cauchy-Schwarz inequality. So, if $\|w\|$ is too small, the risk value may never reach the specified γ.

This theorem describes how the training of F_1 affects the data requirement for training f_2. In particular, a small $t(F_1)$ results in a small n_2 for the right hand side of the inequality to stay fixed at a specific value. This indicates that Step 2 needs the fewest fully-labeled data to produce a solution that attains a particular test performance if Step 1 produced a minimizer for $t(F_1)$. This theorem then confirms that Step 1 indeed produces a minimizer for $t(F_1)$ since it finds an element in $\mathbb{S} \cap \mathbb{D}$ given enough sufficiently-labeled data. Therefore, the more sufficiently-labeled data Step 1 can leverage, the better it can find an element in $\mathbb{S} \cap \mathbb{D}$, and, consequently, the fewer fully-labeled training data Step 2 will need to attain a particular test performance.

The intuition behind why Step 1 can help learn a minimizer of $t(F_1)$ can be explained as follows. Given a data representation, i.e., an F_1, $t(F_1)$ characterizes the model capacity needed from f_2 to minimize the classification risk to a certain degree. And one can solve the problem with the simplest model (smallest $t(F_1)$) when and only when given the easiest (most separable) data representation. Step 1 essentially pursues an F_1 that offers the most separable data representation.

---

2. Any positive risk value can be improved to 0 by the trivial solution $w = 0$. Therefore, only negative risk values are relevant.

Note that this result only goes in one direction. Specifically, the training of f_2 or the amount of fully-labeled examples given does not necessarily affect the data requirement of training F_1. This is intuitive since the learning algorithm trains F_1 and f_2 sequentially and freezes F_1 completely after its training.

2.6 Sufficient Labels Are Easier to Obtain Directly Than Full Labels

Sufficiently-labeled data is less costly to obtain than fully-labeled data in many ways. Determining whether a given pair of examples are from the same class is conceptually simpler for an annotator compared to determining the specific class of each of the examples in the pair. For example, to tell if two car images correspond to the same car is typically much easier than specifying the make and model of car contained in each individual image. Also, for certain sensitive topics, people may be more comfortable with reporting sufficient labels rather than full labels. In the United States, for example, sometimes people are unwilling to disclose whether they voted or who they voted for in elections. Therefore, instead of surveying each person individually and asking them if they voted or who they voted for in the past election, it would perhaps make the surveyees more comfortable if they are paired with their friend and are asked if they made the same or different voting decisions.

Moreover, we can justify that it is simpler to collect sufficiently-labeled data from a learning theoretical perspective. Specifically, we show that learning to produce full labels for unlabeled examples has a larger sample complexity than learning to produce sufficient labels for unlabeled example pairs as the number of underlying classes increases. In other words, training a competent model to assign sufficient labels for example pairs would require fewer labeled training examples than training a competent model assigning full labels.

To see this, the key observation is that the Gaussian complexity of a c-class classification problem using common loss functions such as the hinge loss or cross-entropy loss is $\mathcal{O}\left(c^2/\sqrt{n}\right)$ in general, where n is the labeled training sample size (Mohri et al., 2018). Generating sufficient labels is always a two-class classification problem regardless of the number of actual classes. Therefore, its sample complexity is $\mathcal{O}(1/\sqrt{n})$, whereas the sample complexity for generating full labels is $\mathcal{O}\left(c^2/\sqrt{n}\right)$.[3] This justifies our claim.

Note that this result evidently only applies in the generic multi-class classification setting. Despite the lack of rigorous theoretical justification on our algorithm in the multi-class case, our empirical results demonstrate strong performance from our method in this setting, making our method highly relevant in practice for multi-class classification.

3. A More General Learning Algorithm

In Section 2.2, we proposed a learning method for learning with sufficient labels. The method was proposed for binary classification with the hinge loss. And it used a specific loss function formulation in its Step 1.

---

3. There is an implicit price to pay when learning to generate sufficient labels, which is that the input dimension is doubled. This increases certain terms in typical Gaussian/Rademacher complexity bounds by, e.g., addition or multiplication by a fixed scalar (Bartlett and Mendelson, 2002; Sun et al., 2016). But the increase is constant, which will be eventually dominated by the quadratic growth in c.

In this section, we extend this algorithm to classification with c classes $(c \geq 2)$ using arbitrary loss function. And we present other alternatives that can be used as the loss function in Step 1.

Here, we assume that the model admits the form

$$F = F_2 \circ F_1, F_1 : \mathbb{X} \to \mathbb{R}^p, F_2 : \mathbb{R}^p \to \mathbb{R}^c \tag{29}$$

$$F_2 : u \mapsto (\langle w_1, \phi(u) \rangle + b_1, \ldots, \langle w_c, \phi(u) \rangle + b_c)^\top, \tag{30}$$

where F_2 is a vector of binary classifiers, each classifying one class versus the rest. The ith coordinate of F_2 corresponds to the ith class. At test time, the model outputs the class corresponding to the maximum coordinate of F_2. Clearly, many popular classifier networks including the ResNets (He et al., 2016) admit this representation.

Let two loss functions $\ell : \mathbb{R}^c \times \{1, \ldots, c\} \to \mathbb{R}$, $\ell_1 : \mathbb{R}^p \times \mathbb{R}^p \times \{0, 1\} \to \mathbb{R}$ be given. For example, ℓ can be softmax followed by cross-entropy. Also suppose we have a hypothesis space and two sets of labeled data as in the case of Section 2.2 except that now there are potentially more than two classes in the dataset. The two steps of this more general algorithm are as follows.

i Step 1 (Training the Hidden Layers): Find an $\hat{F}_1$ in

$$\underset{F_1 \in \mathbb{F}_1}{\arg\min} \frac{1}{n_1} \sum_{i=1}^{n_1} \ell_1 \left(F_1 \left(x_{1_i} \right), F_1 \left(x'_{1_i} \right), T \left(y_{1_i}, y'_{1_i} \right) \right). \tag{31}$$

ii Step 2 (Training the Output Layer): Find an $\hat{F}_2$ in

$$\underset{F_2 \in \mathbb{F}_2}{\arg\min} \frac{1}{n_2} \sum_{i=1}^{n_2} \ell \left(F_2 \circ \hat{F}_1 \left(x_{2_i} \right), y_{2_i} \right). \tag{32}$$

iii Return: $\hat{F}_2 \circ \hat{F}_1$.

Clearly, when $c = 2$, implementing F_2 as a real-valued function without bias, choosing ℓ_1 as in Eq. 12, and then letting ℓ be the unbounded binary hinge loss in Section 2.1 recovers the algorithm proposed in Section 2.2.

In addition to the ℓ_1 we used in Section 2.2, there are other alternatives that can learn an F_1 that satisfies the requirements in Theorem 3 when plugged into Eq. 31. Some of these alternatives are based on the "proxy objectives" proposed by Duan et al. (2020).

Define bivariate function $k(u, v) = \langle \phi(u), \phi(v) \rangle$. Let $\beta := \min_{u,v} k(u, v)$ and define $k_i^\star = \mathbb{1} \left(T \left(y_{1_i}, y'_{1_i} \right) = 1 \right) r^2 + \mathbb{1} \left(T \left(y_{1_i}, y'_{1_i} \right) = 0 \right) \beta$. Denote the vector in $\mathbb{R}^{n_1}$ whose ith element is $k \left(F_1 \left(x_{1_i} \right), F_1 \left(x'_{1_i} \right) \right)$ as K_{F_1} and the vector whose ith element is $k_i^\star$ as $K^\star$.

- Negative cosine similarity (NCS)[4]:

$$\ell_1 \left(F_1 \left(x_{1_i} \right), F_1 \left(x'_{1_i} \right), T \left(y_{1_i}, y'_{1_i} \right) \right) = -n_1 \frac{k \left(F_1 \left(x_{1_i} \right), F_1 \left(x'_{1_i} \right) \right) k_i^\star}{\|K_{F_1}\|_2 \|K^\star\|_2} \tag{33}$$

---

4. When plugged into Eq. 31, the empirical risk to be minimized becomes the negated cosine similarity between K_{F_1} and $K^\star$, hence the name. In the kernel learning literature, a loss similar to this is known as kernel alignment (Cristianini et al., 2006). In the information theoretical learning literature, another similar loss is known as Cauchy-Schwarz divergence (Principe et al., 2000).

- Contrastive[5]:

$$\ell_1\left(F_1\left(x_{1_i}\right), F_1\left(x'_{1_i}\right), T\left(y_{1_i}, y'_{1_i}\right)\right) = -\log\left(\frac{\sum_{i=1}^{n_1} \mathbb{1}\left(T\left(y_{1_i}, y'_{1_i}\right) = 1\right) e^{k\left(F_1\left(x_{1_i}\right), F_1\left(x'_{1_i}\right)\right)}}{\sum_{i=1}^{n_1} e^{k\left(F_1\left(x_{1_i}\right), F_1\left(x'_{1_i}\right)\right)}}\right)$$
(34)

- Mean Squared Error:

$$\ell_1\left(F_1\left(x_{1_i}\right), F_1\left(x'_{1_i}\right), T\left(y_{1_i}, y'_{1_i}\right)\right) = \left(k\left(F_1\left(x_{1_i}\right), F_1\left(x'_{1_i}\right)\right) - k_i^\star\right)^2$$
(35)

We do not offer rigorous theoretical justification on these extensions of our algorithm. Nevertheless, their strong performance on challenging large-scale datasets (as we shall demonstrate later) hints that more general versions of our theoretical results exist for these more generic learning settings. For example, our earlier theoretical analysis may be extended to this multi-class classifier model by noting that each of the output node is an one-class-versus-the-rest (binary) classifier. We leave this as future work.

4. Class-Encrypted Learning With Sufficient Labels

In important domains including finance and healthcare, data security and user privacy protection need to (and are sometimes required to) be taken into consideration when developing machine learning solutions (Dwork et al., 2014; Kaissis et al., 2020). In particular, traditional training datasets for supervised learning are especially vulnerable to label information leaks because they associate each individual example directly with its full label, giving away all information the moment the database is breached. Techniques such as federated learning (Konečný et al., 2015, 2016), differential privacy (Dwork et al., 2014), and homomorphic encryption (Acar et al., 2018) have been proposed to address various different aspects of security and privacy. In this section, we discuss how converting full labels into sufficient labels can serve as a generic encryption scheme that is orthogonal and complementary to the established privacy-preserving techniques. Note that the notion of privacy preservation we discuss here is a generic one and should not be confused with the more specific one studied in differential privacy (Dwork et al., 2014).

Sufficiently-labeled data can naturally serve as a layer of encryption on user information that is impossible to fully penetrate using only this given data. This can be easily understood through an example.

Suppose one wants to train a predictor for some disease using a database of real patient records with each labeled with a doctor's diagnosis. Such a label is certainly sensitive information and to protect data security and user privacy, the hospital, where the diagnoses were done and the full results known, can convert the majority of fully-labeled patient data locally to sufficiently-labeled data before transporting and/or storing this data for training. Suppose the hospital follows the process described in Algorithm 1, which ensures that no single individual appears twice in the sufficiently-labeled set and that the patients in the sufficiently-labeled set do not overlap with those in the fully-labeled set. Then,

---

5. This is an extension of a popular loss function from contrastive learning (Saunshi et al., 2019).

Algorithm 1 An example algorithm for converting part of a fully-labeled set of patient records into a sufficiently-labeled one

---

Require: Size of sufficiently-labeled training set $n \geq 1$, fully-labeled patient records $P :=$ $\{(x_i, y_i)\}_{i=1}^N, N > 2n$, with each x_i corresponding to a different individual and y_i his/her diagnosis

    $a \leftarrow 0$
    $Q \leftarrow \emptyset$
    while $a < n$ **do**
        Randomly choose $(x_b, y_b) \in P$ *s.t.* $x_b \neq x_a$
        $Q \leftarrow Q \cup \{(x_a, x_b, T(y_a, y_b))\}$
        $P \leftarrow P \setminus \{(x_a, y_a), (x_b, y_b)\}$
        $a \leftarrow a + 1$
    end while
    return A sufficiently-labeled set Q

---

class information of users in the sufficiently-labeled set is encrypted. Specifically, for any individual in the sufficiently-labeled set, people with access to the data can at most know if this individual has identical result with another random individual, but not the actual diagnosis itself.

As we have shown, this sufficiently-labeled data (the Q returned by Algorithm 1), together with a small subset of unconverted fully-labeled data (the rest of P after performing Algorithm 1), suffices to train performant predictors. The fully-labeled set can be as small as one example per class. This pipeline is illustrated in Fig. 2.

In general, in a sufficiently-labeled dataset $\{(x_i, x_i', T(y_i, y_i'))\}_{i=1}^n$, the actual values of y_i and y_i' cannot be re-identified if only given $T(y_i, y_i')$. This can help protect privacy if y_i represents some sensitive information about user x_i. In this case, only relative information between user pairs is stored in the dataset, and absolution class information of each individual user is effectively encrypted and cannot be recovered even when the dataset is breached. This layer of encryption requires very little extra computational resource, and according to our theory, one can train models on this encrypted data without compromising performance.[6] And it works regardless of the number of classes, contrasting some other forms of privacy-preserving labels that cease to be effective in certain cases such as binary classification (Ishida et al., 2017). Obviously, sufficient labels can be used alongside many existing privacy-protecting techniques for enhanced security.

4.1 Encryption Strength

Given some fully-labeled user data $P := \{(x_i, y_i)\}_{i=1}^N, x_i \neq x_j, \forall i \neq j$, there are many ways of encoding them into a sufficiently-labeled dataset Q,[7] and not all of them have the same encryption strength.

---

6. Our learning method still requires a small set of fully-labeled data that would reveal y_i for each user in this set. But we have shown that this set can be as small as a single randomly-chosen example per class. Protecting the safety of such a small dataset is a much easier task.
7. We assume here that each user x_a from P appears at least once in Q (so that we are not wasting data) and that for any $(x_a, x_b, T(y_a, y_b))$ in Q, we have $a \neq b$.
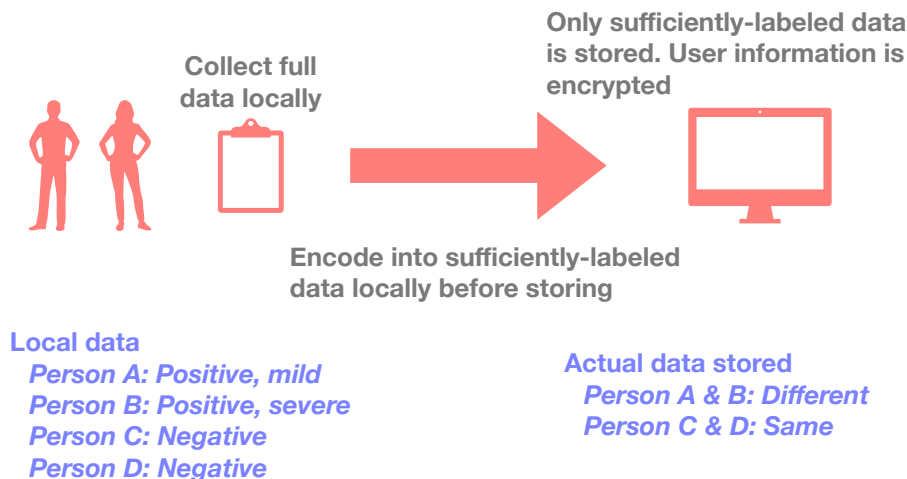
Figure 2: A sufficiently-labeled dataset can serve as a natural layer of encryption on sensitive user information that cannot be fully cracked. In this disease diagnosis example, even if a malicious user gains full access to the sufficiently-labeled dataset, this person can at most see if two randomly-paired individuals have the same result, but not the actual diagnoses themselves. Such a dataset can still be used to train a performant predictor for diagnosing this disease.

At one extreme, each user x_a in P appears $N-1$ times in Q. In other words, we form Q from pairwise combinations of all distinct examples in P. This will give the largest Q (it contains $N \times (N-1)/2$ examples). But the encryption strength is the weakest in the sense that, because pairwise relationships among all examples are available in Q, one can group examples into clusters with each cluster containing all examples from one class and one class only. Nevertheless, the actual class label of each cluster remains unknown. But if some prior knowledge on the class distribution is given, one may succeed in guessing the true label of each cluster.

At the other extreme, each x_a in P appears exactly once in Q (one process that will lead to this case is Algorithm 1). In this case, the converted Q will be the smallest (it contains $\lfloor N/2 \rfloor$ examples). On the other hand, this provides the strongest encryption among all possible Qs if each x_a is paired with a randomly-chosen x_b. Specifically, all that one can deduce about any x_a using this dataset is whether it is from the same class as some other random x_b. And evidently, the more balanced the class distribution is, the less revealing this piece of information about x_a (and x_b) will be.

We leave as future work a rigorous study on the quantitative encryption strength of our method and its connections with other established privacy-preserving techniques.

5. Related Work

5.1 Learning With Easier-to-Obtain Training Data

In classification, designing learning algorithms that can produce performant models using training data that is easier to obtain has been a long-sought goal. Depending on how "easier to obtain" is interpreted, we can classify previous works into the following two categories. Works such as that by Sohn et al. (2020) do not change the underlying learning algorithms, which rely on fully-labeled data, but instead seek to reduce the cost of obtaining labels by synthesizing them instead. On the other hand, Ishida et al. (2017); Saunshi et al. (2019) and others propose novel learning algorithms that can leverage information from other forms of labeled data to reduce need for fully-labeled data. These alternative forms of labeled data are easier to obtain directly. We review these two categories of methods separately.

5.1.1 Reducing the Cost of Obtaining Full Labels

A popular way to reduce the amount of human supervision needed is to synthesize (full) labels on a large set of unlabeled data by leveraging a smaller fully-labeled set and some prior knowledge. The synthesized full labels are then used for training (Sohn et al., 2020; Lee, 2013; Sajjadi et al., 2016b; Laine and Aila, 2017; Berthelot et al., 2019b,a; Miyato et al., 2018; Xie et al., 2019; Tarvainen and Valpola, 2017; Sajjadi et al., 2016a; Xie et al., 2020). Some of the popular approaches to obtain these synthesized labels including using network's predictions on unlabeled data (Lee, 2013; Laine and Aila, 2017), using network's internal randomness such as dropout to implicitly augment labeled data (Sajjadi et al., 2016b), and actively augmenting examples within classes (Sohn et al., 2020; Xie et al., 2019; Berthelot et al., 2019a; Xie et al., 2019; Miyato et al., 2018). State-of-the-art results are often obtained by combining multiple methods for synthesizing labels (Berthelot et al., 2019b; Sohn et al., 2020; Xie et al., 2019; Berthelot et al., 2019a). These methods are sometimes called "semi-supervised" learning (Oliver et al., 2018).

These methods can be seen as ones that reduce the cost for obtaining fully-labeled training data, and are therefore orthogonal to our approach, which aims at learning from a novel form of labeled data (sufficiently-labeled) that is intrinsically easier to obtain than fully-labeled data. Despite the same goal of facilitating learning with easier-to-obtain training data, these methods ultimately rely on the assumption that the unlabeled training set is strongly related to the fully-labeled one, which is fundamentally task-dependent. In fact, this assumption can be easily violated in real-world datasets, in which case the synthesized labels are so noisy that they poison training rather than helping it (Oliver et al., 2018). In contrast, by approaching from a different angle, our method is inherently immune to this issue, and task-agnostic performance bounds have been drawn. As future work, our method may be combined with the above reviewed methods such that the cost of obtaining sufficient labels can be reduced as well.

5.1.2 Leveraging Other Forms of Labeled Data

These works proposed new learning algorithms that can directly learn from labeled data of other forms in addition to fully-labeled data. These novel supervision forms include, for

example, pairwise (dis)similarities, complementary labels, partial labels, and even implicit labels.

Learning from pairwise similarities and dissimilarities. Some works explored learning for classification using data pairs each with a binary label indicating whether this pair of examples are from the same class or not (Bao et al., 2018; Shimada et al., 2021; Shi et al., 2021; Hsu et al., 2019; Bao et al., 2020; Balcan et al., 2008; Bellet et al., 2012). These pairs are called similar or dissimilar pairs in these works, depending on if they are from the same class. Pairwise similarities and dissimilarities are the same as sufficient labels. Therefore, these works are similar to ours.

Bao et al. (2018) derived an unbiased estimator to the true classification risk using terms only involving similar pairs and unlabeled data. They called this learning setting similar-unlabeled (SU) classification. Shi et al. (2021) proposed an improved optimization scheme for SU classification. Shimada et al. (2021) generalized the SU setting to also include dissimilar pairs, leading to the so-called similar-dissimilar-unlabeled classification problem. The authors proposed estimators to the true classification risk that involved only similar, dissimilar, and unlabeled data. These estimators were shown to lead to better error bounds than that obtained from SU classification. In the above works, only mild assumptions were made on the loss function for the theoretical results to hold. Hsu et al. (2019) derived an empirical estimator to a specific classification risk (the negative log-likelihood (NLL)) using terms involving only similar and dissimilar pairs. For a specific binary classification loss function, Bao et al. (2020) established that the classification risk can be reformulated into a risk that only involves sufficiently-labeled data up to label flipping on the classifier. Balcan et al. (2008) defined (ϵ, γ, τ)-goodness, which is a characterization of representations using only sufficiently-labeled data. Then they showed that for (ϵ, γ, τ)-good representations, a linear classifier that optimizes the hinge loss exists with high probability. Below, we discuss how these prior works differ from ours.

Hsu et al. (2019) and Bao et al. (2020) both demonstrated the validity of learning with a mixture of sufficiently-labeled data and fully-labeled data, echoing our work.[8] However, these works are limited in the following regards. First, neither work discussed how the sample size of the sufficiently-labeled set affects that of the fully-labeled set. In contrast, we proved that having more sufficiently-labeled data reduces the need for fully-labeled data. This is important because it shows that we can not only use sufficient labels *with* full labels, but also use sufficient labels *in place of* full labels. In practice, we demonstrated that, with enough sufficiently-labeled data, one only needs as few as one randomly-chosen full label per class to obtain good performance. Successful training with such a minimal need for full labels has not been shown in these prior works, and it further proves the efficacy of learning with sufficient labels. Second, Hsu et al. (2019) did not discuss the sample complexity of learning with sufficiently-labeled data versus that of learning with purely fully-labeled data. While Bao et al. (2020) offered such an analysis, they did not scale their evaluations beyond toy data. Sample complexity is a key characterization of efficient learning, and describes how "informative" for training one sufficiently-labeled datum is compared to a fully-labeled one. Through extensive evaluations, we established that learning from sufficient labels enjoys similar sample complexity as learning from purely full labels even on challenging

---

8. Hsu et al. (2019) required full labels for learning the mapping between output nodes and semantic classes. Bao et al. (2020) proposed to use class priors, which requires fully-labeled data to estimate in practice.

high-dimensional datasets. These are important results that make sufficient labels even more of a compelling alternative to full labels. We also drew connection between sufficient labels and sufficient statistics, offering a new perspective on this form of supervision. Albeit not the main focus on the paper, we pointed out the potential value of sufficient labels as an encryption method.

Bao et al. (2018); Shimada et al. (2021) focused on learning with a combination of sufficiently-labeled and unlabeled data. And they compared only with unsupervised and semi-supervised baselines that learned with the same data. In comparison, we aim at learning with sufficiently-labeled and a small set of fully-labeled data. Further, our proposed learning scheme leads to performance similar to supervised learning using full labels — a much stronger baseline. We think that their assumption of learning without any full label is of theoretical value but somewhat overly restrictive in practice. Even though it is true that obtaining a large set of fully-labeled training data can be time and resource-consuming, obtaining a small set with a few examples per class is usually a trivial task. Bao et al. (2018); Shimada et al. (2021) also only considered binary classification.

(ϵ, γ, τ)-goodness (Balcan et al., 2008; Bellet et al., 2012) and variants (Wang et al., 2009) are characterizations of the optimal representation using only sufficiently-labeled data, parallel to our characterization (Theorem 3). Their optimality guarantee is also similar to ours: That an optimal (in terms of, e.g., hinge loss) downstream linear classifier exists. (ϵ, γ, τ)-goodness framework requires strong constraint on the hypothesis space of F_1, i.e., that it must be formed from a bivariate function (metric) evaluated on some training examples (Theorem 1 in (Bellet et al., 2012)). Practical algorithms such as that proposed by Bellet et al. (2012) also impose parametric assumptions on the form of this bivariate function. In comparison, our theoretical analysis is generic in these regards.

Learning from complementarily-labeled data. A complementary label on example X is a class to which X does not belong. Ishida et al. (2017) proposed a framework to learn from complementarily-labeled data — training sets in which each example is given one complementary label. Upon some assumptions on data distribution and loss function formulation, a family of empirical risks using only complementarily-labeled data were shown to be unbiased estimators to the true classification risk. Therefore, learning to minimize these empirical risks produces solutions that minimize the true risk, given enough data. Ishida et al. (2019) extended this work to arbitrary losses. Feng et al. (2020) tackled the complementary label learning setting where multiple complementary labels were available for each example. All above works need complementarily-labeled data only but can be modified to also utilize fully-labeled data. Yu et al. (2018) generalized the work by Ishida et al. (2017) to the case where, given the true label, the distribution of the complementary labels can be arbitrary rather than uniform. Estimation of a transition matrix requires fully-labeled data.

Complementarily-labeled data is a function of fully-labeled data, which makes it a statistic, similar to our sufficiently-labeled data. A complementary label, like a sufficient label, is also easier to obtain than a full label.

Contrastive learning. Contrastive learning methods leverage implicitly-labeled data for pre-training the network body. Then a network head, usually a single linear layer, is trained on top of the network body using fully-labeled data for some downstream task (Saunshi et al., 2019). The main benefit is that the sample complexity for learning the network

head can be reduced due to contrastive pre-training. Contrastive learning methods are characterized by the use of a contrastive loss on example pairs when training the network body, and have seen great success in natural language processing (Mikolov et al., 2013; Logeswaran and Lee, 2018; Pagliardini et al., 2017) and computer vision (Wang and Gupta, 2015).

The data used in contrastive pre-training is actually unlabeled for the downstream task. But we consider them to have been implicitly labeled since each pair of examples are considered similar (in the same class) or dissimilar (not in the same class) according to some prior knowledge such as co-occurrence, which is a form of (noisy) implicit label on pairs. In fact, theoretical guarantees have only be made when strong connections between these implicit classes and the classes in the downstream task can be assumed (Saunshi et al., 2019).

Contrastive pre-training does not require explicitly-labeled examples with sufficient labels, whereas our learning method does. In terms of the connections between this work and contrastive learning, a typical contrastive loss enforces examples from the same (implicit) class to be mapped closer to each other in the final feature space and examples from different classes to be mapped far apart (Saunshi et al., 2019). And this is exactly what our learning algorithm attempts to do when training the network body, and the optimality of this objective in our setting is guaranteed by our Theorem 3. Thus, our learning framework can be viewed as a supervised learning analog of contrastive learning. Further, the fact that having more sufficient labels reduces the need for full labels when training the output layer echos the result that contrastive learning reduces the labeled sample complexity of the downstream task (Saunshi et al., 2019).

Learning under imperfect supervision. Many existing works consider situations where supervision is imperfect. In practice, slightly imperfect supervision can be much less costly to obtain, making these methods highly relevant. One example is when only a specific class can be annotated effectively. Another is when the labels are noisy.

Partial label learning addresses the learning settings where multiple labels are available for each instance, but only one of them is correct (Cour et al., 2011; Feng and An, 2018, 2019a,b; Zhang and Yu, 2015). Noisy label learning assumes that there is a certain degree of error in the labels, which is common in practice (Han et al., 2018b,a; Menon et al., 2015; Wei et al., 2020; Xia et al., 2019). Learning from positive unlabeled data is to learn from only examples from a specific class and unlabeled data in a binary classification setting (du Plessis et al., 2014, 2015; Elkan and Noto, 2008; Kiryo et al., 2017; Sakai et al., 2017, 2018). This is relevant in practice, for example, when the cost for annotating a specific class is prohibitively high. Lu et al. (2019) considered binary classification and proposed a setting in which the (unlabeled) training data is sampled from two marginals with complementary class priors (requires labeled data to approximate in practice). Specifically, the first set of X should be sampled from a marginal distribution with $p(Y = +1) = \theta$ and the second set from a marginal with $p(Y = +1) = \theta' \neq \theta$. Then they showed that, if the true class prior $p(Y = +1)$ of the test distribution is known, one can derive an unbiased risk estimator using only these two sets of training data. This setting can be viewed as sampling two fully-labeled training sets with noisy labels, with the first set of examples from class $Y = +1$ having $1 - \theta$ probability of incorrect labels, and the other set of examples from class $Y = -1$

having θ' incorrect label probability. In practice, the best performance is obtained with θ being close to 1 and θ' close to 0, i.e., when the noise level is low.

We address the problem of learning from a new source of supervision that we call sufficient labels. If we understand the notion of imperfect supervision to be one that includes any information source that is weaker and less complete than full labels, then we can regard sufficient labels as a type of imperfect supervision since they are summaries (less comprehensive) of full labels.

5.2 Modular Training of Deep Architectures

Our proposed learning algorithms are modular in the sense that the underlying network is viewed as two components that are trained separately. Similar provably optimal modular learning algorithms have been proposed by, e.g., Duan et al. (2019, 2020). These methods also used a form of labeled training data similar to our sufficiently-labeled data. However, these works analyzed the learning algorithms from the perspective of modular training, not the training data used. In contrast, this work gives a complete analysis on the efficacy of learning with sufficient labels. And, unlike these works, our theoretical results are not based on the assumption that the sufficiently-labeled data for training the hidden layers are derived from the fully-labeled data for training the output layer. This enables optimality guarantees in settings where the sufficiently-labeled data are obtained directly from annotators without collecting the more expensive fully-labeled data first.

6. Experiments: Training With (Almost Only) Sufficiently-Labeled Data

In this section, we verify our first two claims in Section 2.3 through experiments. Specifically, we test for both the hinge loss and the cross-entropy loss on MNIST, Fashion-MNIST (Xiao et al., 2017), SVHN (Netzer et al., 2011), CIFAR-10, and CIFAR-100 (Krizhevsky and Hinton, 2009) that

1. we can train state-of-the-art classifiers with a mixture of sufficiently-labeled and fully-labeled data, and

2. only a small set of fully-labeled data is needed if given enough sufficiently-labeled data.

In particular, we show that the reduction on the need for fully-labeled data in practice is even more drastic that what is guaranteed by Theorem 4: A single randomly-chosen full label per class was all the learning algorithm needed in all cases tested. This means that state-of-the-art classifiers can be trained with almost only sufficiently-labeled data in practice.

6.1 The Basic Settings

The settings for these experiments are as follows. We trained a LeNet-5 (LeCun et al., 1998) for MNIST and a ResNet-18 (He et al., 2016) for Fashion-MNIST, SVHN, and CIFAR-10. To make sure that the model satisfied the assumption in Section 2.1, the activation vector of the nonlinearity before the final linear layer was always normalized to 1 by (elementwise)

Table 1: Mean accuracy ± standard deviation from 5 trials. Best result in bold. For "Data Usage", "Full" refers to the number of fully-labeled examples, and "Suff." the number of sufficiently-labeled examples. For the models that used 10/100 fully-labeled examples, these examples were randomly picked at each trial, with one example from each class. Strong classifiers can be learned from almost entirely sufficiently-labeled data.

Dataset (Model Used)	Loss	Data Usage		Test Acc. (%)
		Full	Suff.	
MNIST (LeNet-5)	Hinge	60k	0	99.22 ± 0.05
		10	120k	98.97 ± 0.15
		10	240k	99.12 ± 0.08
		10	online	$\mathbf{99.23 \pm 0.05}$
	Cross-Entropy	60k	0	$\mathbf{99.32 \pm 0.08}$
		10	120k	98.98 ± 0.15
		10	240k	99.12 ± 0.07
		10	online	99.23 ± 0.05
Fashion-MNIST (ResNet-18)	Hinge	60k	0	$\mathbf{95.14 \pm 0.12}$
		10	120k	93.85 ± 0.29
		10	240k	94.61 ± 0.17
		10	online	95.03 ± 0.28
	Cross-Entropy	60k	0	$\mathbf{95.10 \pm 0.18}$
		10	120k	93.89 ± 0.32
		10	240k	94.63 ± 0.15
		10	online	95.03 ± 0.27
SVHN (ResNet-18)	Hinge	$73,257$	0	96.49 ± 0.07
		10	$146,514$	95.77 ± 0.12
		10	$293,028$	96.11 ± 0.11
		10	online	$\mathbf{96.66 \pm 0.16}$
	Cross-Entropy	$73,257$	0	96.46 ± 0.10
		10	$146,514$	95.78 ± 0.11
		10	$293,028$	96.13 ± 0.10
		10	online	$\mathbf{96.67 \pm 0.16}$
CIFAR-10 (ResNet-18)	Hinge	50k	0	94.09 ± 0.21
		10	100k	90.99 ± 0.33
		10	200k	93.56 ± 0.24
		10	online	$\mathbf{94.24 \pm 0.25}$
	Cross-Entropy	50k	0	94.19 ± 0.13
		10	100k	91.00 ± 0.34
		10	200k	93.63 ± 0.17
		10	online	$\mathbf{94.24 \pm 0.24}$
CIFAR-100 (ResNet-18)	Hinge	50k	0	73.22 ± 0.36
		100	400k	70.45 ± 0.17
		100	800k	72.67 ± 0.32
		100	online	$\mathbf{73.79 \pm 0.46}$
	Cross-Entropy	50k	0	$\mathbf{74.18 \pm 0.15}$
		100	400k	70.58 ± 0.24
		100	800k	72.67 ± 0.28
		100	online	73.98 ± 0.50

dividing itself by its norm. This normalization did not affect performance. The optimizer used was stochastic gradient descent with batch size 128. Each hidden module (or full model, in the cases where only full labels were used) was trained with step size 0.1, 0.01, and 0.001, for 200, 100, and 50 epochs, respectively. For CIFAR-100 in "online" mode, the hidden module was trained for 1000, 800, and 400 epochs with the said step sizes. If needed, each output layer was trained with step size 0.1 for 50 epochs. For MNIST, the data was preprocessed by training set sample mean subtraction and then division by training set sample standard deviation. For Fashion-MNIST, SVHN, CIFAR-10, and CIFAR-100, the data was randomly cropped and flipped after the said normalization procedure. For a training set of size $50k/100k/200k/60k/120k/240k/73,257/146,514/293,028/400k/800k$, we randomly selected $5k/10k/20k/5k/10k/20k/5k/10k/20k/40k/80k$ training examples to form the validation set. The validation set was used for tuning hyperparameters and for determining the best model to save during each training session. For a training set of size 10, no validation data was used and the optimal model was chosen based on the convergence of the loss function value on training data. For SVHN, we did not use the "additional" training images. For all datasets, we report performance on the standard test sets. And standard supervised learning (training using only and all of the available fully-labeled data), which is the strongest baseline there is, will be used for comparison. We did not notice any of the proposed ℓ_1's significantly outperforming the others. In all cases, we report performance using NCS.

6.2 Obtaining Sufficient Labels

All of these curated datasets we selected are fully-labeled, and the sufficiently-labeled datasets used for the following experiments were derived from fully-labeled data. For example, to create a sufficiently-labeled training set with size 100k, we randomly sampled 100k pairs of examples from the original fully-labeled training set. And for each pair, a sufficient label of 1 or 0 was generated based on their original full labels. We discarded some pairs that do not contain useful information. Specifically, suppose the input examples in the original dataset are $\{x_i\}_{i=1}^n$, then the sampled set of example pairs is $\{x_{1_i}, x_{2_i}\}_{i=1}^{n_1}$ with $1_i < 2_i$ and $(1_i, 2_i) \neq (1_j, 2_j), \forall i \neq j$. Such a sufficiently-labeled training set covers $n_1/(n(n-1)/2)$ of all possible informative example pairs from the original training set.

To approximate the performance upper bound for training with sufficiently-labeled data using these curated datasets, we used an online random sampling strategy ("online"). Namely, in each training epoch, we iterated the original fully-labeled dataset in batches of size 128 and for each batch of fully-labeled data, we converted it into a set of sufficiently-labeled pairs (by taking pairwise combination, with uninformative pairs discarded as before) and used them to compute a model update step. And since the dataset was randomly shuffled at each epoch, we can potentially exhaust all sufficiently-labeled pairs given enough training epochs. In "online", the size of the validation set was chosen based on the size of the original fully-labeled dataset using the rules above.

6.3 Results

From Table 1, we see that state-of-the-art classifiers can be trained using almost only sufficiently-labeled data. Indeed, a single randomly-chosen fully-labeled example from each

class — an amount that is trivial to obtain in practice — sufficed when learning with sufficient labels in all cases. This validates Theorem 3 and Theorem 4. Also, to achieve similar test performance, learning with a mixture of sufficiently-labeled and fully-labeled data requires a training set with size similar to that of the training set required by learning with only fully-labeled data. This verifies the earlier claim that learning with sufficient labels enjoys similar sample complexity as learning with purely full labels.

Note that the sufficiently-labeled datasets we used were converted from datasets curated for the purpose of standard supervised training with fully-labeled data, and these "repurposed" datasets can be suboptimal for learning with sufficient labels due to the number of different-class pairs significantly outnumbering that of the same-class pairs. Take CIFAR-100 as an example, the fully-labeled dataset has 100 classes with the data population carefully chosen such that each class is represented by precisely 1% of the population, then if we convert the dataset into a sufficiently-labeled one by exhausting all pairwise combinations, examples with sufficient label 0 will dominate the resulting sufficiently-labeled population (roughly 99% sufficient labels will be 0). In online mode, we found that good performance can still be obtained after prolonged training, which is perhaps because the model will eventually see enough same-class pairs in terms of absolute example count. In the modes where we trained with a predetermined number of sufficiently-labeled examples, we controlled the ratio of same-class to different-class pairs to make sure that there are enough same-class pairs in the training set. Specifically, when we created the sufficiently-labeled training set by sampling examples from the original fully-labeled set, the sampling was performed in batches. In each batch, the sampler randomly selected M unique classes, where M was much smaller than the total number of classes. Then, for this batch, it randomly sampled data pairs only from these M classes. The need for keeping enough same-class pairs in the dataset is perhaps why we needed more data to match the performance of online or training with fully-labeled data especially on datasets with many classes such as CIFAR-100.

In general, a class distribution calibrated for fully-labeled learning may not be ideal for sufficiently-labeled learning. And we expect the performance of learning with sufficient labels to improve with tailored datasets.

7. Conclusion

We proposed a novel form of labeled training data for classification that we called "sufficiently-labeled" data. Sufficiently-labeled data can be obtained directly from annotators and it is easier to collect sufficiently-labeled data than fully-labeled data. Sufficiently-labeled data can also serve as a layer of encryption on user information that cannot be fully broken.

A training algorithm that can learn from a mixture of sufficiently-labeled and fully-labeled data was proposed and analyzed. We proved that it can learn performant models and empirically showed that it shares a similar sample complexity as ERM using purely fully-labeled data. We also showed that having more sufficient labels means needing less full labels. And, in practice, we demonstrated that state-of-the-art classifiers can be trained with as few as a single random full label from each class when given enough sufficient labels.

Acknowledgments

Appendix A. Proof of Theorem 3

Proof

$$R(f_2 \circ F_1, X, Y) \tag{36}$$

$$= E_{X,Y}\ell(f_2 \circ F_1(X), Y)$$

$$= E_Y E_{X|Y}\ell(f_2 \circ F_1(X), Y) \tag{37}$$

$$= \alpha E_{X|Y=+}\ell_+(f_2 \circ F_1(X)) + (1-\alpha)E_{X|Y=-}\ell_-(f_2 \circ F_1(X)) \tag{38}$$

$$= \alpha \int \ell_+(f_2 \circ F_1(x))p_{X|Y=+}(x)dx + (1-\alpha) \int \ell_-(f_2 \circ F_1(x))p_{X|Y=-}(x)dx \tag{39}$$

$$= \int \int \left[\alpha\ell_+(f_2 \circ F_1(x)) + (1-\alpha)\ell_-(f_2 \circ F_1(x'))\right] p_{X|Y=+}(x)p_{X'|Y'=-}(x')dxdx' \tag{40}$$

$$= \int \int \left[\alpha\ell_+(f_2 \circ F_1(x)) + (1-\alpha)\ell_-(f_2 \circ F_1(x'))\right] p_{X,X'|Y=+,Y'=-}(x, x')dxdx' \tag{41}$$

$$= E_{X,X'|Y=+,Y'=-} \left[\alpha\ell_+(f_2 \circ F_1(X)) + (1-\alpha)\ell_-(f_2 \circ F_1(X'))\right], \tag{42}$$

where $\ell_+(u) = -u$ and $\ell_-(u) = +u$, and we have used the conditional independence of X, X'.

$$R(f_2 \circ F_1, X, Y) \tag{43}$$

$$= E_{X,X'|Y=+,Y'=-} \left[-\left\langle w, \alpha\phi \circ F_1(X) - (1-\alpha)\phi \circ F_1(X')\right\rangle\right]$$

$$= E_{X,X'|Y=+,Y'=-} \left[\|w\| \cos\theta(w, F_1, X, X') \left(-\|\alpha\phi \circ F_1(X) - (1-\alpha)\phi \circ F_1(X')\|\right)\right], \tag{44}$$

where $\theta(w, F_1, X, X')$ is the angle between w and $\alpha\phi \circ F_1(X) - (1-\alpha)\phi \circ F_1(X')$ defined through the inner product when

$$\|w\| > 0 \text{ and } \|\alpha\phi \circ F_1(X) - (1-\alpha)\phi \circ F_1(X')\| > 0 \tag{45}$$

and can be assigned any arbitrary value when either norm is 0 since this leaves the expectation unchanged.

Note that $F_1 \in \mathbb{S}$ if and only if

$$\Pr\left(\phi \circ F_1(X) = u_+|Y = +\right) = \Pr\left(\phi \circ F_1(X') = u_-|Y' = -\right) = 1, \tag{46}$$

for some constant vectors $u_+ \neq \lambda u_-, \forall\lambda \geq 0$. The "if" direction is obvious. To prove the other direction, note that by conditional independence, if $u \in \text{supp}(\phi \circ F_1(X))$ given $Y = +$ (or $-$) and $v \in \text{supp}(\phi \circ F_1(X'))$ given $Y' = Y$, then $(u, v) \in \text{supp}((\phi \circ F_1(X), \phi \circ F_1(X')))$ given $Y = Y'$. Now, suppose u_1, u_2 with $u_1 \neq u_2$ are both in $\text{supp}(\phi \circ F_1(X))$ given $Y = +$ (or $-$) and let $v \in \text{supp}(\phi \circ F_1(X'))$ given $Y = Y'$, then $\phi \circ F_1(X) = \phi \circ F_1(X')$ cannot be true w.p. 1 given $Y = Y'$ since both (u_1, v) and (u_2, v) are in the support and v cannot be equal to u_1 and u_2 simultaneously. Therefore, $\phi \circ F_1(X) = \phi \circ F_1(X')$ w.p. 1 given $Y = Y'$ implies that $\phi \circ F_1(X)$ is degenerate given Y. To prove $u_+ \neq \lambda u_-, \forall\lambda \geq 0$, note that since $\|u_+\| = \|u_-\| > 0$, the only possible λ is 1. But if $u_+ = u_-$, then given $Y \neq Y'$, $\phi \circ F_1(X) = \phi \circ F_1(X')$ w.p. 1, a contradiction.

Now, to prove

$$\mathbb{F}_1^\star \supseteq \mathbb{D} \cap \mathbb{S}, \tag{47}$$

it suffices to show that if

$$F_1^\star \in \arg\min_{F_1 \in \mathbb{F}_1} E_{X,X'|Y=+,Y'=-} \left(-\left\| \phi \circ F_1(X) - \phi \circ F_1(X') \right\|^2 \right) \tag{48}$$

$$= \arg\min_{F_1 \in \mathbb{F}_1} E_{X,X'|Y=+,Y'=-} \left(\langle \phi \circ F_1(X), \phi \circ F_1(X') \rangle \right), \text{ and}$$

$$\Pr\left(\phi \circ F_1^\star(X) = u_+^\star | Y = + \right) = \Pr\left(\phi \circ F_1^\star(X') = u_-^\star | Y' = - \right) = 1, \tag{49}$$

for some $u_+^\star \neq \lambda u_-^\star, \forall \lambda \geq 0$, then

$$\min_{f_2 \in \mathbb{F}_2} R(f_2 \circ F_1^\star, X, Y) \leq \min_{f_2 \in \mathbb{F}_2} R(f_2 \circ F_1, X, Y), \forall F_1 \in \mathbb{F}_1. \tag{50}$$

To this end, note that Eq. 49 implies $\alpha\phi \circ F_1^\star(X) - (1-\alpha)\phi \circ F_1^\star(X')$ is a nonzero constant vector w.p. 1 given $Y = +, Y' = -$. Therefore, given $Y = +, Y' = -$, for any $1/r \geq a > 0$, there exists $w^\star$ with $\|w^\star\| = a$ and $\cos\theta(w^\star, F_1^\star, X, X') = 1$ w.p. 1.[9]

For any $F_1 \in \mathbb{F}_1$, let an $f_2 \in \arg\min_{f_2 \in \mathbb{F}_2} R(f_2 \circ F_1, X, Y)$ be parameterized by w. If $\|w\| > 0$, we can then find $w^\star$ such that $\|w^\star\| = \|w\|$ and $\cos\theta(w^\star, F_1^\star, X, X') = 1$ w.p. 1. On the other hand, if $\|w\| = 0$, find $w^\star$ such that $\|w^\star\| = a$ and $\cos\theta(w^\star, F_1^\star, X, X') = 1$ w.p. 1 for some $1/r \geq a > 0$. In both cases, we have

$$\min_{f_2 \in \mathbb{F}_2} R(f_2 \circ F_1, X, Y) \tag{51}$$

$$= E_{X,X'|Y=+,Y'=-} \left[\|w\| \cos\theta(w, F_1, X, X') \left(-\left\| \alpha\phi \circ F_1(X) - (1-\alpha)\phi \circ F_1(X') \right\| \right) \right]$$

$$\geq E_{X,X'|Y=+,Y'=-} \left[\|w^\star\| \left(-\left\| \alpha\phi \circ F_1(X) - (1-\alpha)\phi \circ F_1(X') \right\| \right) \right] \tag{52}$$

$$= \|w^\star\| E_{X,X'|Y=+,Y'=-} \left(-\left\| \alpha\phi \circ F_1(X) - (1-\alpha)\phi \circ F_1(X') \right\| \right) \tag{53}$$

$$= \|w^\star\| E_{X,X'|Y=+,Y'=-} \left(-\sqrt{\left\| \alpha\phi \circ F_1(X) - (1-\alpha)\phi \circ F_1(X') \right\|^2} \right) \tag{54}$$

$$\overset{\text{(Jensen's Inequality)}}{\geq} \|w^\star\| \left(-\sqrt{E_{X,X'|Y=+,Y'=-} \left\| \alpha\phi \circ F_1(X) - (1-\alpha)\phi \circ F_1(X') \right\|^2} \right) \tag{55}$$

$$= \|w^\star\| \left(-\sqrt{\left(\alpha^2 + (1-\alpha)^2 \right) r^2 - 2\alpha(1-\alpha) E_{X,X'|Y=+,Y'=-} \langle \phi \circ F_1(X), \phi \circ F_1(X') \rangle} \right) \tag{56}$$

$$\overset{\text{(Eq. 48)}}{\geq} \|w^\star\| \left(-\sqrt{\left(\alpha^2 + (1-\alpha)^2 \right) r^2 - 2\alpha(1-\alpha) E_{X,X'|Y=+,Y'=-} \langle \phi \circ F_1^\star(X), \phi \circ F_1^\star(X') \rangle} \right) \tag{57}$$

$$= \|w^\star\| \left(-\sqrt{E_{X,X'|Y=+,Y'=-} \left\| \alpha\phi \circ F_1^\star(X) - (1-\alpha)\phi \circ F_1^\star(X') \right\|^2} \right) \tag{58}$$

$$\overset{\text{(Eq. 49)}}{=} \|w^\star\| \left(-\sqrt{\left\| \alpha u_+^\star - (1-\alpha)u_-^\star \right\|^2} \right) \tag{59}$$

$$= \|w^\star\| E_{X,X'|Y=+,Y'=-} \left(-\left\| \alpha\phi \circ F_1^\star(X) - (1-\alpha)\phi \circ F_1^\star(X') \right\| \right) \tag{60}$$

$$= E_{X,X'|Y=+,Y'=-} \left[\|w^\star\| \cos\theta(w^\star, F_1^\star, X, X') \left(-\left\| \alpha\phi \circ F_1^\star(X) - (1-\alpha)\phi \circ F_1^\star(X') \right\| \right) \right] \tag{61}$$

$$\geq \min_{f_2 \in \mathbb{F}_2} R(f_2 \circ F_1^\star, X, Y). \tag{62}$$

---

9. In fact, by Cauchy-Schwarz inequality, $a \left(\alpha u_+^\star - (1-\alpha)u_-^\star \right) / \left\| \alpha u_+^\star - (1-\alpha)u_-^\star \right\|$ is the unique $w^\star$.

On the other hand, to prove

$$\mathbb{F}_1^\star \subseteq \mathbb{S} \cap \mathbb{D}, \tag{63}$$

let an $F_1 \in \mathbb{F}_1^\star$ be given, we need to show

$$F_1 \notin \mathbb{S}^{\mathsf{c}} \cup \mathbb{D}^{\mathsf{c}} = (\mathbb{S}^{\mathsf{c}} \cap \mathbb{D}) \cup \mathbb{D}^{\mathsf{c}}. \tag{64}$$

If $F_1 \in \mathbb{D}^{\mathsf{c}}$, there exists $F_1^\star \in \mathbb{S}$ such that

$$E_{X,X'|Y=+,Y'=-} \left(- \left\| \phi \circ F_1(X) - \phi \circ F_1(X') \right\|^2 \right) \tag{65}$$
$$> E_{X,X'|Y=+,Y'=-} \left(- \left\| \phi \circ F_1^\star(X) - \phi \circ F_1^\star(X') \right\|^2 \right).$$

Let an $f_2 \in \arg\min_{f_2 \in \mathbb{F}_2} R(f_2 \circ F_1, X, Y)$ be parameterized by w. We have shown that we can find $w^\star$ such that $\|w^\star\| = \|w\|$ (or $\|w^\star\| = a$ for some $1/r \geq a > 0$ if $\|w\| = 0$) and $\cos\theta(w^\star, F_1^\star, X, X') = 1$ w.p. 1. Then we have

$$\min_{f_2 \in \mathbb{F}_2} R(f_2 \circ F_1, X, Y) \tag{66}$$
$$= E_{X,X'|Y=+,Y'=-} \left[\|w\| \cos\theta(w, F_1, X, X') \left(- \left\| \alpha\phi \circ F_1(X) - (1-\alpha)\phi \circ F_1(X') \right\| \right) \right]$$
$$\geq E_{X,X'|Y=+,Y'=-} \left[\|w^\star\| \left(- \left\| \alpha\phi \circ F_1(X) - (1-\alpha)\phi \circ F_1(X') \right\| \right) \right] \tag{67}$$
$$= \|w^\star\| E_{X,X'|Y=+,Y'=-} \left(- \left\| \alpha\phi \circ F_1(X) - (1-\alpha)\phi \circ F_1(X') \right\| \right) \tag{68}$$
$$\overset{\text{(Similar to Eq. 53 - Eq. 60)}}{>} \|w^\star\| E_{X,X'|Y=+,Y'=-} \left(- \left\| \alpha\phi \circ F_1^\star(X) - (1-\alpha)\phi \circ F_1^\star(X') \right\| \right) \tag{69}$$
$$= E_{X,X'|Y=+,Y'=-} \left[\|w^\star\| \cos\theta(w^\star, F_1^\star, X, X') \left(- \left\| \alpha\phi \circ F_1^\star(X) - (1-\alpha)\phi \circ F_1^\star(X') \right\| \right) \right] \tag{70}$$
$$\geq \min_{f_2 \in \mathbb{F}_2} R(f_2 \circ F_1^\star, X, Y), \tag{71}$$

contradicting $F_1 \in \mathbb{F}_1^\star$.

On the other hand, if $F_1 \in \mathbb{S}^{\mathsf{c}} \cap \mathbb{D}$, then let $F_1^\star \in \mathbb{S} \cap \mathbb{D}$ and we have

$$E_{X,X'|Y=+,Y'=-} \left(- \left\| \phi \circ F_1(X) - \phi \circ F_1(X') \right\|^2 \right) \tag{72}$$
$$= E_{X,X'|Y=+,Y'=-} \left(- \left\| \phi \circ F_1^\star(X) - \phi \circ F_1^\star(X') \right\|^2 \right).$$

Let an $f_2 \in \arg\min_{f_2 \in \mathbb{F}_2} R(f_2 \circ F_1, X, Y)$ be parameterized by w. Again, we can find $w^\star$ such that $\|w^\star\| = \|w\|$ (or $\|w^\star\| = a$ for some $1/r \geq a > 0$ if $\|w\| = 0$) and $\cos\theta(w^\star, F_1^\star, X, X') = 1$ w.p. 1.

Suppose we can show that there does not exist w such that w is linearly dependent with $\alpha\phi \circ F_1(X) - (1-\alpha)\phi \circ F_1(X')$ w.p. 1 given $Y = +, Y' = -$, $\|\alpha\phi \circ F_1(X) - (1-\alpha)\phi \circ F_1(X')\| >$

0, then

$$\min_{f_2 \in \mathbb{F}_2} R(f_2 \circ F_1, X, Y) \tag{73}$$

$$= E_{X,X'|Y=+,Y'=-} \left[-\langle w, \alpha\phi \circ F_1(X) - (1-\alpha)\phi \circ F_1(X') \rangle \right]$$

$$> E_{X,X'|Y=+,Y'=-} \left[\|w\| \left(-\|\alpha\phi \circ F_1(X) - (1-\alpha)\phi \circ F_1(X')\| \right) \right] \tag{74}$$

$$= \|w^\star\| E_{X,X'|Y=+,Y'=-} \left(-\|\alpha\phi \circ F_1(X) - (1-\alpha)\phi \circ F_1(X')\| \right) \tag{75}$$

$$\overset{\text{(Similar to Eq. 53 - Eq. 60)}}{\geq} \|w^\star\| E_{X,X'|Y=+,Y'=-} \left(-\|\alpha\phi \circ F_1^\star(X) - (1-\alpha)\phi \circ F_1^\star(X')\| \right) \tag{76}$$

$$= E_{X,X'|Y=+,Y'=-} \left[\|w^\star\| \cos\theta(w^\star, F_1^\star, X, X') \left(-\|\alpha\phi \circ F_1^\star(X) - (1-\alpha)\phi \circ F_1^\star(X')\| \right) \right] \tag{77}$$

$$\geq \min_{f_2 \in \mathbb{F}_2} R(f_2 \circ F_1^\star, X, Y), \tag{78}$$

which, again, contradicts $F_1 \in \mathbb{F}_1^\star$.

Finally, to show that there does not exist w such that w is linearly dependent with $\alpha\phi \circ F_1(X) - (1-\alpha)\phi \circ F_1(X')$ w.p. 1 given $Y = +, Y' = -, \|\alpha\phi \circ F_1(X) - (1-\alpha)\phi \circ F_1(X')\| > 0$, we use the fact that $F_1 \notin \mathbb{S}$. First, note that for any w, $\alpha\phi \circ F_1(X) - (1-\alpha)\phi \circ F_1(X')$ being linearly dependent with w w.p. 1 given $Y = +, Y' = -, \|\alpha\phi \circ F_1(X) - (1-\alpha)\phi \circ F_1(X')\| > 0$ implies that the support of $\alpha\phi \circ F_1(X) - (1-\alpha)\phi \circ F_1(X')$ given $Y = +, Y' = -$ is a subset of $\{\lambda w | \lambda \in \mathbb{R}\}$. By conditional independence of X, X' and the assumption that $\|\phi(u)\| = r, \forall u$, any $u \in \text{supp}(\alpha\phi \circ F_1(X))$ given $Y = +$ and $v \in \text{supp}((1-\alpha)\phi \circ F_1(X'))$ given $Y' = -$ have $u - v \in \text{supp}((\alpha\phi \circ F_1(X) - (1-\alpha)\phi \circ F_1(X')))$ given $Y = +, Y' = -$ and $\|u\| = \alpha/(1-\alpha)\|v\|$. Meanwhile, for a given v

$$u - v = \lambda w \Rightarrow \|v + \lambda w\| = \|u\| = \frac{\alpha}{1-\alpha}\|v\| \tag{79}$$

can be true for at most two distinct values of λ. This implies

$$|\text{supp}(\phi \circ F_1(X)) \text{ given } Y = +| \leq 2. \tag{80}$$

Similarly, we have

$$|\text{supp}(\phi \circ F_1(X')) \text{ given } Y' = -| \leq 2. \tag{81}$$

Eq. 80 and 81 together gives

$$|\text{supp}(\phi \circ F_1(X))| \leq 4, \tag{82}$$

contradicting $F_1 \notin \mathbb{S}$ due to the assumption that $|\text{supp}(\phi \circ F_1(X))| > 4, \forall F_1 \notin \mathbb{S}$. ∎

Appendix B. Proof of Theorem 4

Proof For any F_1, define

$$\ell \circ \mathbb{F}_{2,A} \circ F_1 = \{\ell \circ f_2 \circ F_1 | f_2 \in \mathbb{F}_{2,A}\}. \tag{83}$$

Then let $\mathcal{R}_n$ denote the Rademacher complexity, we have (Bartlett and Mendelson, 2002),

$$\mathcal{R}_n(\ell \circ \mathbb{F}_{2,A} \circ F_1) \leq \frac{Ar}{\sqrt{n}}. \tag{84}$$

Note that

$$|\ell(f_2 \circ F_1(x), y)| \leq \|w\| r \leq Ar, \forall f_2 \in \mathbb{F}_{2,A}. \tag{85}$$

Let a true risk value $\gamma < 0$ and any probability δ be given, then, with probability at least $1 - \delta$, we have (Bartlett and Mendelson, 2002; Shalev-Shwartz and Ben-David, 2014)

$$\sup_{\hat{f}_2 \in \hat{\mathbb{F}}_{2,F_1,\gamma}^{\star}} R\left(\hat{f}_2 \circ F_1, X, Y\right) - \gamma \leq 2\frac{t\left(F_1\right) r}{\sqrt{n_2}} + 5t\left(F_1\right) r \sqrt{\frac{2 \ln\left(8/\delta\right)}{n_2}}, \tag{86}$$

where $t(F_1)$ is the function that gives the smallest A such that for all $f_2 \in \mathbb{F}_{2,F_1,\gamma}$, we have $\|w\| \leq A$.

Now, to prove $\mathbb{S} \cap \mathbb{D} \subseteq \arg \min_{F_1 \in \mathbb{F}_1} t(F_1)$, let $F_1^{\star} \in \mathbb{S} \cap \mathbb{D}$ and an arbitrary eligible F_1 be given. We have seen in the proof of Theorem 3 that $\|\alpha\phi \circ F_1^{\star}(X) - (1-\alpha)\phi \circ F_1^{\star}(X')\| \neq 0$ w.p. 1 given $Y = +, Y' = -$. Therefore, $E_{X,X'|Y=+,Y'=-} \left(- \|\alpha\phi \circ F_1^{\star}(X) - (1-\alpha)\phi \circ F_1^{\star}(X')\|\right) \neq 0$. Also, define $\theta\left(w, F_1, X, X'\right)$ as in the proof of Theorem 3. Let a particular f_2 be the f_2 with the smallest $\|w\|$ among all f_2 such that $f_2 \circ F_1$ attains true risk γ. Suppose this f_2 is parameterized by w. Evidently, $t(F_1) = \|w\|$. And we have

$$\gamma = R(f_2 \circ F_1, X, Y) \tag{87}$$

$$= E_{X,X'|Y=+,Y'=-} \|w\| \cos\theta(w, F_1, X, X') \left(- \|\alpha\phi \circ F_1(X) - (1-\alpha)\phi \circ F_1(X')\|\right) \tag{88}$$

$$\geq \|w\| E_{X,X'|Y=+,Y'=-} \left(- \|\alpha\phi \circ F_1(X) - (1-\alpha)\phi \circ F_1(X')\|\right) \tag{89}$$

$$\overset{\text{(Jensen's Inequality)}}{\geq} \|w\| \left(-\sqrt{E_{X,X'|Y=+,Y'=-} \|\alpha\phi \circ F_1(X) - (1-\alpha)\phi \circ F_1(X')\|^2}\right) \tag{90}$$

$$= \|w\| \sqrt{\frac{E_{X,X'|Y=+,Y'=-} \|\alpha\phi \circ F_1(X) - (1-\alpha)\phi \circ F_1(X')\|^2}{E_{X,X'|Y=+,Y'=-} \|\alpha\phi \circ F_1^{\star}(X) - (1-\alpha)\phi \circ F_1^{\star}(X')\|^2}} \tag{91}$$

$$\left(-\sqrt{E_{X,X'|Y=+,Y'=-} \|\alpha\phi \circ F_1^{\star}(X) - (1-\alpha)\phi \circ F_1^{\star}(X')\|^2}\right). \tag{92}$$

Using an argument from the proof of Theorem 3 (paragraph following Eq. 50), $F_1^{\star} \in \mathbb{S}$ implies the existence of an $w^{\star}$ such that

$$\|w^{\star}\| = \|w\| \sqrt{\frac{E_{X,X'|Y=+,Y'=-} \|\alpha\phi \circ F_1(X) - (1-\alpha)\phi \circ F_1(X')\|^2}{E_{X,X'|Y=+,Y'=-} \|\alpha\phi \circ F_1^{\star}(X) - (1-\alpha)\phi \circ F_1^{\star}(X')\|^2}} \leq \|w\| \tag{93}$$

and $\cos\theta(w^{\star}, F_1^{\star}, X, X') = 1$ w.p. 1, where the inequality is due to Eq. 48. Then, with derivation same as Eq. 55 - Eq. 60, we have

$$\gamma \geq \|w^{\star}\| E_{X,X'|Y=+,Y'=-} \cos\theta(w^{\star}, F_1^{\star}, X, X') \left(- \|\alpha\phi \circ F_1^{\star}(X) - (1-\alpha)\phi \circ F_1^{\star}(X')\|\right) \tag{94}$$

$$\geq \min_{f_2 \in \mathbb{F}_{2,\|w^{\star}\|}} R(f_2 \circ F_1^{\star}, X, Y), \tag{95}$$

which, combined with $F_1^\star \in \mathbb{D}$ and the fact that another $w^{\star\star}$ can be found (by, e.g., multiplying $w^\star$ with a scalar) such that it parameterizes an $f_2^{\star\star}$ with $f_2^{\star\star} \circ F_1^\star$ attaining risk γ and $\|w^{\star\star}\| \leq \|w^\star\|$, indicates

$$t(F_1^\star) \leq \|w^\star\| \leq \|w\| = t(F_1), \tag{96}$$

proving $\mathbb{S} \cap \mathbb{D} \subseteq \arg\min_{F_1 \in \mathbb{F}_1} t(F_1)$.

On the other hand, assume $F_1 \notin \mathbb{S} \cap \mathbb{D}$ but $F_1 \in \arg\min_{F_1 \in \mathbb{F}_1} t(F_1)$. Suppose a particular f_2 is the f_2 with the smallest $\|w\|$ among all f_2 such that $f_2 \circ F_1$ attains γ true risk. Let this particular f_2 be parameterized by w and we have $t(F_1) = \|w\|$. Let $F_1^\star \in \mathbb{S} \cap \mathbb{D}$ be given.

If $F_1 \notin \mathbb{D}$, then using a similar argument as above, we can find $w^\star$ with $\|w^\star\| < \|w\|$ such that $f_2^\star \circ F_1^\star$ attains at most γ true risk and therefore $t(F_1^\star) \leq \|w^\star\| < \|w\| = t(F_1)$, contradicting $F_1 \in \arg\min_{F_1 \in \mathbb{F}_1} t(F_1)$.

If $F_1 \in \mathbb{D}$ but $F_1 \notin \mathbb{S}$, then we have shown in the proof of Theorem 3 that there exists $w^\star$ such that $\|w^\star\| = \|w\|, \cos\theta(w^\star, F_1^\star, X, X') = 1$, but $R(f_2^\star \circ F_1^\star, X, Y) < R(f_2 \circ F_1, X, Y) = \gamma$, indicating that there exists $w^{\star\star}$ parameterizing an $f_2^{\star\star}$ with $R(f_2^{\star\star} \circ F_1^\star, X, Y) = \gamma$ but $\|w^{\star\star}\| < \|w^\star\| = \|w\|$ (one such $w^{\star\star}$ is given by $\omega w^\star$ for some $\omega \in [0, 1)$). Thus, $t(F_1^\star) \leq \|w^{\star\star}\| < \|w\| = t(F_1)$, again contradicting $F_1 \in \arg\min_{F_1 \in \mathbb{F}_1} t(F_1)$. ∎

References

Abbas Acar, Hidayet Aksu, A Selcuk Uluagac, and Mauro Conti. A survey on homomorphic encryption schemes: Theory and implementation. *ACM Computing Surveys (CSUR)*, 51 (4):1–35, 2018.

Maria-Florina Balcan, Avrim Blum, and Nathan Srebro. Improved guarantees for learning via similarity functions. In *21st Annual Conference on Learning Theory (COLT 2008)*, pages 287–298, 2008.

Han Bao, Gang Niu, and Masashi Sugiyama. Classification from pairwise similarity and unlabeled data. In *International Conference on Machine Learning*, pages 452–461. PMLR, 2018.

Han Bao, Takuya Shimada, Liyuan Xu, Issei Sato, and Masashi Sugiyama. Similarity-based classification: Connecting similarity learning to binary classification. *arXiv preprint arXiv:2006.06207*, 2020.

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

Aurélien Bellet, Amaury Habrard, and Marc Sebban. Similarity learning for provably accurate sparse linear classification. In *International Conference on Machine Learning*, 2012.

David Berthelot, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. Remixmatch: Semi-supervised learning with distribution matching and augmentation anchoring. In *International Conference on Learning Representations*, 2019a.

David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *Advances in Neural Information Processing Systems*, pages 5049–5059, 2019b.

George Casella and Roger L Berger. *Statistical inference*, volume 2. Duxbury Pacific Grove, CA, 2002.

Timothee Cour, Ben Sapp, and Ben Taskar. Learning from partial labels. *The Journal of Machine Learning Research*, 12:1501–1536, 2011.

Nello Cristianini, Jaz Kandola, Andre Elisseeff, and John Shawe-Taylor. On kernel target alignment. In *Innovations in machine learning*, pages 205–256. Springer, 2006.

Marthinus du Plessis, Gang Niu, and Masashi Sugiyama. Convex formulation for learning from positive and unlabeled data. In *International conference on machine learning*, pages 1386–1394, 2015.

Marthinus C du Plessis, Gang Niu, and Masashi Sugiyama. Analysis of learning from positive and unlabeled data. In *Advances in neural information processing systems*, pages 703–711, 2014.

Shiyu Duan, Shujian Yu, Yunmei Chen, and Jose C Principe. On kernel method–based connectionist models and supervised deep learning without backpropagation. *Neural computation*, 32(1):97–135, 2019.

Shiyu Duan, Shujian Yu, and Jose Principe. Modularizing deep learning via pairwise learning with kernels. *arXiv preprint arXiv:2005.05541*, 2020.

Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014.

Charles Elkan and Keith Noto. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 213–220, 2008.

Lei Feng and Bo An. Leveraging latent label distributions for partial label learning. In *IJCAI*, pages 2107–2113, 2018.

Lei Feng and Bo An. Partial label learning with self-guided retraining. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3542–3549, 2019a.

Lei Feng and Bo An. Partial label learning by semantic difference maximization. In *IJCAI*, pages 2294–2300, 2019b.

Lei Feng, Takuo Kaneko, Bo Han, Gang Niu, Bo An, and Masashi Sugiyama. Learning with multiple complementary labels. In *International Conference on Machine Learning*, pages 3072–3081. PMLR, 2020.

Bo Han, Jiangchao Yao, Gang Niu, Mingyuan Zhou, Ivor Tsang, Ya Zhang, and Masashi Sugiyama. Masking: A new perspective of noisy supervision. In *Advances in Neural Information Processing Systems*, pages 5836–5846, 2018a.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in neural information processing systems*, pages 8527–8537, 2018b.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Yen-Chang Hsu, Zhaoyang Lv, Joel Schlosser, Phillip Odom, and Zsolt Kira. Multi-class classification without multi-class labels. In *International Conference on Learning Representations*, 2019.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

Takashi Ishida, Gang Niu, Weihua Hu, and Masashi Sugiyama. Learning from complementary labels. In *Advances in neural information processing systems*, pages 5639–5649, 2017.

Takashi Ishida, Gang Niu, Aditya Menon, and Masashi Sugiyama. Complementary-label learning for arbitrary losses and models. In *International Conference on Machine Learning*, pages 2971–2980. PMLR, 2019.

Georgios A Kaissis, Marcus R Makowski, Daniel Rückert, and Rickmer F Braren. Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence*, pages 1–7, 2020.

Ryuichi Kiryo, Gang Niu, Marthinus C du Plessis, and Masashi Sugiyama. Positive-unlabeled learning with non-negative risk estimator. In *Advances in neural information processing systems*, pages 1675–1685, 2017.

Jakub Konečnỳ, Brendan McMahan, and Daniel Ramage. Federated optimization: Distributed optimization beyond the datacenter. *arXiv preprint arXiv:1511.03575*, 2015.

Jakub Konečnỳ, H Brendan McMahan, Daniel Ramage, and Peter Richtárik. Federated optimization: Distributed machine learning for on-device intelligence. *arXiv preprint arXiv:1610.02527*, 2016.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. 2009.

Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. In *ICLR (Poster)*. OpenReview.net, 2017.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, 2013.

Lajanugen Logeswaran and Honglak Lee. An efficient framework for learning sentence representations. *arXiv preprint arXiv:1803.02893*, 2018.

Nan Lu, Gang Niu, Aditya Krishna Menon, and Masashi Sugiyama. On the minimal supervision for training any binary classifier from only unlabeled data. In *ICLR (Poster)*, 2019.

Aditya Menon, Brendan Van Rooyen, Cheng Soon Ong, and Bob Williamson. Learning from corrupted binary labels via class-probability estimation. In *International Conference on Machine Learning*, pages 125–134, 2015.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

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

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.

Avital Oliver, Augustus Odena, Colin A Raffel, Ekin Dogus Cubuk, and Ian Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. In *Advances in neural information processing systems*, pages 3235–3246, 2018.

Matteo Pagliardini, Prakhar Gupta, and Martin Jaggi. Unsupervised learning of sentence embeddings using compositional n-gram features. *arXiv preprint arXiv:1703.02507*, 2017.

Jose C Principe, Dongxin Xu, John Fisher, and Simon Haykin. Information theoretic learning. *Unsupervised adaptive filtering*, 1:265–319, 2000.

Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Mutual exclusivity loss for semi-supervised deep learning. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 1908–1912. IEEE, 2016a.

Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. In *Advances in neural information processing systems*, pages 1163–1171, 2016b.

Tomoya Sakai, Marthinus Christoffel Plessis, Gang Niu, and Masashi Sugiyama. Semi-supervised classification based on classification from positive and unlabeled data. In *International conference on machine learning*, pages 2998–3006. PMLR, 2017.

Tomoya Sakai, Gang Niu, and Masashi Sugiyama. Semi-supervised auc optimization based on positive-unlabeled learning. *Machine Learning*, 107(4):767–794, 2018.

Nikunj Saunshi, Orestis Plevrakis, Sanjeev Arora, Mikhail Khodak, and Hrishikesh Khandeparkar. A theoretical analysis of contrastive unsupervised representation learning. In *International Conference on Machine Learning*, pages 5628–5637. PMLR, 2019.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Wanli Shi, Bin Gu, Xiang Li, Cheng Deng, and Heng Huang. Triply stochastic gradient method for large-scale nonlinear similar unlabeled classification. *Machine Learning*, 110 (8):2005–2033, 2021.

Takuya Shimada, Han Bao, Issei Sato, and Masashi Sugiyama. Classification from pairwise similarities/dissimilarities and unlabeled data via empirical risk minimization. *Neural Computation*, 33(5):1234–1268, 2021.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv preprint arXiv:2001.07685*, 2020.

Shizhao Sun, Wei Chen, Liwei Wang, Xiaoguang Liu, and Tie-Yan Liu. On the depth of deep neural networks: a theoretical view. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pages 2066–2072, 2016.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in neural information processing systems*, pages 1195–1204, 2017.

Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 2013.

Liwei Wang, Masashi Sugiyama, Cheng Yang, Kohei Hatano, and Jufu Feng. Theory and algorithm for learning with dissimilarity functions. *Neural computation*, 21(5):1459–1484, 2009.

Xiaolong Wang and Abhinav Gupta. Unsupervised learning of visual representations using videos. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802, 2015.

Hongxin Wei, Lei Feng, Xiangyu Chen, and Bo An. Combating noisy labels by agreement: A joint training method with co-regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13726–13735, 2020.

Xiaobo Xia, Tongliang Liu, Nannan Wang, Bo Han, Chen Gong, Gang Niu, and Masashi Sugiyama. Are anchor points really indispensable in label-noise learning? In *Advances in Neural Information Processing Systems*, pages 6838–6849, 2019.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. Unsupervised data augmentation for consistency training. *arXiv preprint arXiv:1904.12848*, 2019.

Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10687–10698, 2020.

Xiyu Yu, Tongliang Liu, Mingming Gong, and Dacheng Tao. Learning with biased complementary labels. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 68–83, 2018.

Min-Ling Zhang and Fei Yu. Solving the partial label learning problem: An instance-based approach. In *IJCAI*, pages 4048–4054, 2015.