Electronic Journal of Statistics
Vol. 10 (2016) 2780–2824
ISSN: 1935-7524
DOI: 10.1214/16-EJS1193

Classification with asymmetric label noise: Consistency and maximal denoising

Gilles Blanchard

Intitut für Mathematik
Universität Potsdam
e-mail: gilles.blanchard@math.uni-potsdam.de

Marek Flaska

Mechanical and Nuclear Engineering
Pennsylvania State University
e-mail: MXF5309@psu.edu

Gregory Handy

Department of Mathematics
University of Utah
e-mail: handy@math.utah.edu

Sara Pozzi

Nuclear Engineering and Radiological Sciences
University of Michigan, Ann Arbor
e-mail: pozzisa@umich.edu

and

Clayton Scott

Electrical and Computer Engineering, Statistics
University of Michigan, Ann Arbor
e-mail: clayscot@umich.edu

Abstract: In many real-world classification problems, the labels of training examples are randomly corrupted. Most previous theoretical work on classification with label noise assumes that the two classes are separable, that the label noise is independent of the true class label, or that the noise proportions for each class are known. In this work, we give conditions that are necessary and sufficient for the true class-conditional distributions to be identifiable. These conditions are weaker than those analyzed previously, and allow for the classes to be nonseparable and the noise levels to be asymmetric and unknown. The conditions essentially state that a majority of the observed labels are correct and that the true class-conditional distributions are "mutually irreducible," a concept we introduce that limits the similarity of the two distributions. For any label noise problem, there is a unique pair of true class-conditional distributions satisfying the proposed conditions, and we argue that this pair corresponds in a certain sense to maximal denoising of the observed distributions.

Our results are facilitated by a connection to "mixture proportion estimation," which is the problem of estimating the maximal proportion of one distribution that is present in another. We establish a novel rate of convergence result for mixture proportion estimation, and apply this to obtain consistency of a discrimination rule based on surrogate loss minimization. Experimental results on benchmark data and a nuclear particle classification problem demonstrate the efficacy of our approach.

MSC 2010 subject classifications: Primary 62H30; secondary 68T10.
Keywords and phrases: Classification, label noise, mixture proportion estimation, surrogate loss, consistency.

Received August 2015.

1. Introduction

In binary classification, one observes multiple realizations of two different classes,

$$X_0^1, \ldots, X_0^m \overset{iid}{\sim} P_0,$$
$$X_1^1, \ldots, X_1^n \overset{iid}{\sim} P_1,$$

where P_0 and P_1, the class-conditional distributions, are probability distributions on a Borel space $(\mathcal{X}, \mathfrak{S})$. The feature vector $X_y^i \in \mathcal{X}$ denotes the i-th realization from class $y \in \{0, 1\}$. The general goal is to construct a classifier from this data.

There are several kinds of noise that can affect a classification problem. A first type of noise occurs when P_0 and P_1 have overlapping supports, meaning that the label is not a deterministic function of the feature vector. In this situation, even an optimal classifier makes mistakes. In this work, we consider a second type of noise, *label noise*, that can occur *in addition to* the first type of noise. With label noise, some of the labels of the training examples are corrupted. We focus in particular on *random* label noise, as opposed to feature-dependent or adversarial label noise.

To model label noise, we represent the training data via contamination models:

$$X_0^1, \ldots, X_0^m \overset{iid}{\sim} \tilde{P}_0 := (1 - \pi_0)P_0 + \pi_0 P_1, \tag{1}$$
$$X_1^1, \ldots, X_1^n \overset{iid}{\sim} \tilde{P}_1 := (1 - \pi_1)P_1 + \pi_1 P_0. \tag{2}$$

According to these mixture representations, each "apparent" class-conditional distribution is in fact a contaminated version of the true class-conditional distribution, where the contamination comes from the other class. Thus, $\tilde{P}_0$ governs the training data with apparent class label 0. A proportion $1 - \pi_0$ of these examples have 0 as their true label, while the remaining π_0 have a true label of 1. Similar remarks apply to $\tilde{P}_1$. The noise is asymmetric in that π_0 need not equal π_1. We emphasize that π_0 and π_1 are unknown. The distributions P_0 and P_1 are also unknown, and we do not wish to impose models for them. In particular,

the supports of P_0 and P_1 may overlap, so that the classes are not separable. Previous work on classification with random label noise, reviewed below, has not considered the problem in this generality.

Our first contribution is to introduce necessary and sufficient conditions on the elements P_0, P_1, π_0, π_1 of the contamination models such that these elements are uniquely determined given $\tilde{P}_0$ and $\tilde{P}_1$. These conditions are the following:

- (Total noise level) $\pi_0 + \pi_1 < 1$,
- (Mutual irreducibility) It is not possible to write P_0 as a nontrivial mixture of P_1 and some other distribution, and *vice versa*.

To shed some light on these conditions, we remark that in the absence of any assumption, the solution (P_0, P_1, π_0, π_1) to (1)–(2), when the contaminated distributions $\tilde{P}_0, \tilde{P}_1$ are given (i.e., in the limit of infinite sample sizes, or population version of the problem), is non-unique. For example, were the condition on total label noise not required, for any solution, swapping the role of classes 0 and 1 would also be a solution (with complementary contamination probabilities), while leaving the apparent labels unchanged.

Furthermore, we describe in detail (at the population level) the geometry of the set of all possible solutions (P_0, P_1, π_0, π_1) to (1)–(2). We argue that for any pair $\tilde{P}_0 \neq \tilde{P}_1$, there always exists a *unique* solution satisfying the above two conditions. Moreover, this solution uniquely corresponds to the maximum possible total label noise level $(\pi_1 + \pi_0)$ compatible with the observed contaminated distributions, and also to the maximum possible total variation separation $\|P_1 - P_0\|_{TV}$ under the condition $\pi_1 + \pi_0 < 1$. In this sense, P_0 and P_1 satisfying the second condition are *maximally denoised* versions of the contaminated distributions.

Our second contribution is to develop a discrimination rule that is universally consistent in the sense that for any $\tilde{P}_0, \tilde{P}_1$, it consistently estimates the optimal classification performance as defined with respect to the maximally denoised distributions (which are the underlying uncontaminated distributions under the above conditions). A key aspect of our contribution is that the label noise proportions π_0 and π_1 are unknown, in contrast to previous work, and the linchpin of our solution is a method for accurately estimating π_0 and π_1. We argue that these proportions can be estimated using methods for *mixture proportion estimation* (MPE), which is the problem of estimating the mixing proportion of one distribution in another. We review previous work on MPE and also establish a new rate of convergence result for MPE that is employed in our analysis.

As a third contribution, we present experimental results indicating that the proposed methodology is practically viable. In particular, we show that π_0 and π_1 can be accurately estimated using the same principles guiding our theory. To illustrate this point, we examine some standard benchmark data sets as well as a real data set from a nuclear particle classification problem that is naturally described by our label noise model.

Portions of this work appeared earlier in Scott et al. [41] and Scott [40]. This longer version integrates those versions and extends them by establishing the necessity of the proposed conditions, a consistency analysis featuring clip-

pable losses, a connection to class probability estimation, and a more thorough literature review.

1.1. Motivating application

This work is motivated by a nuclear particle classification problem that is critical for nuclear nonproliferation and nuclear safeguards. An organic scintillation detector is a device commonly used to detect high-energy neutrons. When a particle interacts with the detector, the energy deposited by the particle is converted to a pulse-shaped voltage waveform, which is then digitally sampled to obtain a feature vector $X \in \mathbb{R}^d$, where d is the number of digital samples. The energy distribution of detected neutrons is characteristic of the nuclear source material, and these energy distributions can be inferred from the heights of the observed pulses. However, these detectors are also sensitive to gamma rays, which are frequently emitted by the same fission events that produce neutrons, and which are also strongly present in background radiation. Therefore, to render organic scintillation detectors useful for characterization of nuclear materials, it is necessary to classify between neutron and gamma-ray pulses, a problem referred to as pulse shape discrimination (PSD) [1, 3].

Unfortunately, even in controlled laboratory settings, it is very difficult to obtain pure samples of neutron and gamma-ray pulses. As previously mentioned, the fission events that produce neutrons also yield gamma rays, and gamma rays also arrive from background radiation. Although pure gamma-ray sources do exist, when collecting measurements from such sources, neutrons from the background cannot be completely eliminated. If we view gamma-rays as class 0, by taking a strong and pure gamma-ray source, π_0 will be small but nonzero. On the other hand, the proportion of gamma-rays emitted during fission is intrinsic to the source material, and cannot be changed. Thus π_1 could be in the neighborhood of one-half. With additional time-of-flight information, this proportion can be reduced, but is still non-negligible [3]. Thus, PSD is naturally described by the proposed label noise model. We study this problem empirically in Section 11.

1.2. Label flipping model for label noise

Random label noise can also be modeled according to the label flipping probabilities

$$\mu_i := \Pr(\tilde{Y} = 1 - i \,|\, Y = i).$$

In the label flipping model, a "clean" training data set is corrupted by flipping the labels according to μ_0 and μ_1, independent of X. If we assume Y and $\tilde{Y}$ are jointly distributed, then π_i and μ_i are related via Bayes' rule. The preferred perspective for a given label noise problem, contamination or label flipping, is application dependent. For example, the contamination model better suits the nuclear particle classification problem described above. We also find it more natural to discuss identifiability in terms of the contamination model.

1.3. Related work

Classification in the presence of label noise has drawn the attention of numerous researchers [17]. One common approach is to assume that corrupted labels are more likely to be associated with outlying data points. This has inspired methods to clean, correct, or reweight the training data [10, 34], as well as the use of robust (usually nonconvex) losses [29, 45, 28, 16, 14]. The above approaches are not necessarily based on a random label noise model, but rather assume that noisy labels are more common near the decision boundary.

Generative models have also been applied in the context of random label noise. These impose parametric models on the data-generating distributions, and include the label noise as part of the model. The parameters are then estimated using an EM algorithm [9]. The method of [23] employs kernels in this approach, allowing for the modeling of more flexible distributions.

Negative results for convex risk minimization in the presence of label noise have been established by Long and Servido [26] and Manwani and Sastry [27]. These works demonstrate a lack of noise tolerance for boosting and empirical risk minimization based on convex losses, and suggest that any approach based on convex risk minimization will require modification of the loss, such that the risk minimizer is the optimal classifier with respect to the uncontaminated distributions. Along these lines, Stempfel and Ralaivola [44], Natarajan et al. [32] recently developed such algorithms based on convex losses. The works, however, assume knowledge of the label noise proportions. In the sequel, we establish a consistent discrimination rule that does not assume knowledge of π_0 and π_1; in fact the main focus of the present work is on the estimation of those quantities.

Recently Yang et al. [46] established performance guarantees for multiple kernel learning with noisy labels. This work does not assume label noise is independent of the feature vector, but does require knowledge of the total amount of label noise.

Classification with random label noise has also been studied in the PAC literature. Most PAC formulations assume that (i) P_0 and P_1 have non-overlapping support (i.e., there is a deterministic "target concept" that provides the true labels), (ii) the label noise is symmetric (i.e., independent of the true class label), and (iii) the performance measure is the probability of error [4, 20, 5, 13, 11, 19]. Under these conditions, it typically suffices to train on the contaminated data; only the sample complexity changes. The case of asymmetric label noise was addressed by Blum and Mitchell [8] under condition (i), as the basis of co-training. Some new directions and a thorough review of this body of work were recently presented in [18]. As we discuss in the next section, new challenges emerge when conditions (i), (ii), and (iii) are not assumed.

To our knowledge, previous work under the asymmetric noise model has not addressed a minimal set of conditions for either consistent classification or for consistent estimation of the label noise proportions.

Classification with label noise is related to several other machine learning problems. When $\pi_1 = 0$, we have "one-sided" label noise, and the problem reduces to learning from positive and unlabeled examples (LPUE), also known

as semi-supervised novelty detection (SSND); see Blanchard et al. [7] for a review of this literature. In particular, Blanchard et al. [7] develop theory for "mixture proportion estimation" that we leverage in our analysis.

A basic version of multiple instance learning can be reduced to classification with one-sided label noise [see 36]. In multiple instance learning, the learner is presented with bags of instances. In one basic setting, the bags are labeled negative if they contain only negative instances, and positive if they contain at least one positive instance. If one assumes that the instances in positive bags follow a mixture model $\tilde{P}_1 = (1-\pi)P_1 + \pi P_0$, and the instances are iid according to P_0 or $\tilde{P}_1$, the setting is that of one-sided label noise.

As mentioned above, classification with label noise is the basis of co-training [8], which is a framework for classifying instances that are represented by two distinct "views." The original analysis of co-training considers the "realizable" case, where labels are a deterministic function of inputs. Our results allow us to state a result for co-training without making this restrictive assumption. This result is presented in Section 8.

There is also a connection between classification with label noise and class probability estimation. As pointed out in our initial technical report [42], there is a simple way to express mutual irreducibility in terms of the class probability function. From this relationship, and given other developments in this paper, it is straightforward to express π_0 and π_1 in terms of the maximum and minimum values of the contaminated class probability function. This suggests an alternative estimation strategy for the label noise proportions, which has recently been investigated by Liu and Tao [25] and Menon et al. [30]. In Section 9, we elaborate on this approach and connections to these works. We also investigate this approach experimentally in Section 11.

As a final connection with existing literature, we note that an alternative way to view the contamination model (1)–(2) is to interpret it as a *source separation* problem. In the usual source separation setting, the *realizations* from the different sources are linearly mixed, whereas in the present model, the *source probability distributions* are (we do not observe a signal superposition, but a signal coming randomly from one or the other source). As a common point with the source separation setting, it is necessary to postulate additional constraints on the sources in order to resolve non-uniqueness of the possible solutions. In independent component analysis, for instance, sources are assumed to be independent. Our assumption of mutual irreducibility between the sources plays a conceptually comparable role here. Similarly, the assumption on the total noise level resolves the ambiguity that the sources would otherwise only be identifiable up to permutation.

1.4. Some initial notation

Let $f : \mathcal{X} \to \{0, 1\}$ be a classifier. Denote the (uncontaminated) Type I and Type II errors

$$R_0(f) := P_0(f(X) = 1)$$

$$R_1(f) := P_1(f(X) = 0).$$

These quantities are what define many classification performance measures of interest, such as the so-called *minmax* criterion, $R(f) = \max\{R_0(f), R_1(f)\}$, or the probability of error, $R(f) = \nu R_1(f) + (1-\nu)R_0(f)$, where ν is the a priori probability of class 1.

We also define the corresponding contaminated Type I and II errors

$$\begin{aligned}\tilde{R}_0(f) &:= \tilde{P}_0(f(X) = 1) \\ &= (1-\pi_0)R_0(f) + \pi_0(1 - R_1(f)) \qquad (3)\\ \tilde{R}_1(f) &:= \tilde{P}_1(f(X) = 0) \\ &= (1-\pi_1)R_1(f) + \pi_1(1 - R_0(f)). \qquad (4)\end{aligned}$$

These quantities can easily be estimated from the training data by their basic empirical counterparts.

1.5. Outline

The remainder of the paper is outlined as follows. Section 2 discusses the challenges posed by label noise for classifier design. Section 3 presents an alternate representation of the contamination models that reduces the problem to that of mixture proportion estimation, which is discussed in Section 4. In Section 5 we introduce our proposed identifiability conditions, establish their sufficiency and necessity, and also discuss maximal denoising. A method for mixture proportion estimation is discussed in Section 6, where a novel rate of convergence result is presented and subsequently applied to develop a consistent discrimination rule in Section 7. In Section 8, we apply our label noise results to generalize an earlier result on co-training. Section 9 makes a connection between our label noise framework and the problem of class probability estimation. Algorithm implementations are described in Section 10, and experimental results are provided in Section 11. Shorter proofs tend to be found in the body of the paper, while longer ones are in an appendix.

2. The challenge of label noise

Before delving into more technical matters, we first offer an overview of the challenges posed by label noise. We focus on the population setting ($n_0, n_1 = \infty$) and compare classifier design based on the contaminated distributions, $\tilde{P}_0$ and $\tilde{P}_1$, versus the true ones, P_0 and P_1. To begin, we introduce the following condition on the total amount of label noise.

(A) $\pi_0 + \pi_1 < 1$.

This condition states, in a certain sense, that a majority of the labels are correct on average. It even allows that one of the proportions be very close to one if the other proportion is small enough. This condition was previously adopted by [8].

Let p_0 and p_1 be densities of P_0 and P_1, respectively, with respect to a common dominating measure. Then

$$\tilde{p}_0(x) := (1-\pi_0)p_0(x) + \pi_0 p_1(x),$$
$$\tilde{p}_1(x) := (1-\pi_1)p_1(x) + \pi_1 p_0(x),$$

are respective densities of $\tilde{P}_0$ and $\tilde{P}_1$.

Proposition 1. *Assume* **(A)** *holds. For all* $\gamma \geq 0$*, and every* x *such that* $p_0(x) > 0$ *and* $\tilde{p}_0(x) > 0$*,*

$$\frac{p_1(x)}{p_0(x)} > \gamma \iff \frac{\tilde{p}_1(x)}{\tilde{p}_0(x)} > \lambda,$$

where

$$\lambda = \frac{\pi_1 + \gamma(1-\pi_1)}{1-\pi_0+\gamma\pi_0}. \tag{5}$$

The proof involves a sequence of simple algebraic steps to transform one likelihood ratio into the other, and the use of **(A)** to ensure that the direction of the inequality is preserved.

For most performance measures of interest (probability of error, Neyman-Pearson, etc.), it is well-known that the optimal classifier takes the form of a likelihood ratio test (LRT) based on the true densities [24, 21]. According to the proposition, every true LRT is identical to a contaminated LRT with a different threshold. As the threshold of one LRT sweeps over its range, so too does the threshold of the other LRT. Equivalently, both LRTs generate the same receiver operating characteristic (ROC).

However, if we design a classifier with respect to the contaminated estimates of performance, we will not obtain a classifier that is optimal with respect to the true performance measure, except in very special circumstances. To make this point concrete, we now consider four specific performance measures.

Probability of error. When the feature vector X and label Y are jointly distributed, the probability of misclassification is minimized by a LRT, where the threshold γ is given by the ratio of *a priori* class probabilities. If $\gamma = 1$, then the corresponding threshold for the contaminated LRT is also 1, regardless of π_0 and π_1, which follows directly from (5). Furthermore, assuming $\pi_0, \pi_1 > 0$ and with some simple algebra it is easy to show that $\lambda = \gamma$ only if $\gamma = 1$. Thus, treating the contaminated data as if it were clean is suboptimal whenever the a priori class probabilities are unequal.

Neyman-Pearson. As noted above, the true and contaminated LRTs have the same ROC. If a point on this ROC is chosen such that $\tilde{R}_0(f) = \alpha$, it will generally not be the case that $R_0(f) = \alpha$. This follows because $\tilde{R}_0(f) = (1-\pi_0)R_0(f) + \pi_0 R_1(f)$. Simple algebra shows that $R_0(f) = \tilde{R}_0(f)$ iff $\pi_0 = 0$ or $R_0(f) + R_1(f) = 1$. The latter condition is not satisfied by an optimal classifier unless $P_0 = P_1$, since it corresponds to random guessing. The former case, $\pi_0 = 0$, means the negative class has no contamination, and is equivalent (after swapping class labels) to learning from positive and unlabeled examples.

Minmax. The minmax criterion is defined as $R(f) := \max\{R_0(f), R_1(f)\}$, and the minmax classifier is the minimizer of this quantity. The minmax classifier corresponds to the point on the ROC of the true and contaminated LRTs where $R_0(f) = R_1(f)$. Indeed, if $R_0(f) \neq R_1(f)$, then $\max\{R_0(f), R_1(f)\}$ can be decreased by moving along the ROC such that the larger of $R_0(f), R_1(f)$ is decreased. Thus, designing a classifier with respect to the contaminated distributions yields a point on the optimal ROC where $\tilde{R}_0(f) = \tilde{R}_1(f)$. Using equations (3) and (4), simple algebra reveals that $\tilde{R}_0(f) = \tilde{R}_1(f)$ and $R_0(f) = R_1(f)$ for the same f iff $\pi_0 = \pi_1$ or $R_0(f) = R_1(f) = \frac{1}{2}$. The first condition is not satisfied for asymmetric label noise, and the latter condition is not true for an optimal classifier unless $P_0 = P_1$.

Balanced Error. Menon et al. [30] actually show that the balanced error, given by $\frac{1}{2}(R_0(f) + R_1(f))$, is the only performance measure that is a function of $R_0(f)$ and $R_1(f)$, such that optimizing the corrupted performance measure is equivalent to optimizing the clean performance measure regardless of the label noise proportions or prior class probabilities.

In summary, a classifier that is optimal with respect to a contaminated performance measure is not optimal for the uncontaminated performance measure except in special cases. Accurate estimation of the true performance measure is thus a critical issue for classifier design. In the next section, we expose a technique for estimating performance using estimates of the label noise proportions.

3. Alternate contamination model

We introduce an alternate contamination model that will later be used to obtain estimates of the label noise proportions, and consequently estimates of classifier performance.

Lemma 2. *If $P_0 \neq P_1$ and* **(A)** *holds, then $\tilde{P}_1 \neq \tilde{P}_0$, and there exist unique $0 \leq \tilde{\pi}_0, \tilde{\pi}_1 < 1$ such that*

$$\tilde{P}_0 = (1 - \tilde{\pi}_0)P_0 + \tilde{\pi}_0\tilde{P}_1 \tag{6}$$

$$\tilde{P}_1 = (1 - \tilde{\pi}_1)P_1 + \tilde{\pi}_1\tilde{P}_0. \tag{7}$$

In particular $\tilde{\pi}_0 = \frac{\pi_0}{1-\pi_1} < 1$ and $\tilde{\pi}_1 = \frac{\pi_1}{1-\pi_0} < 1$.

Proof. To see that $\tilde{P}_1 \neq \tilde{P}_0$, assume that equality holds. Plugging in (1)–(2), we obtain

$$(1 - \pi_1 - \pi_0)P_1 = (1 - \pi_1 - \pi_0)P_0,$$

which, since $P_0 \neq P_1$, would imply $\pi_1 + \pi_0 = 1$ and contradict **(A)**.

We turn to identity (6). Matching distributions, the identity holds iff

$$\begin{aligned} P_1(\pi_0 - \tilde{\pi}_0(1 - \pi_1)) &= P_0(1 - \tilde{\pi}_0 + \pi_1\tilde{\pi}_0 - (1 - \pi_0)) \\ &= P_0(\pi_0 - \tilde{\pi}_0(1 - \pi_1)). \end{aligned}$$

Since $P_0 \neq P_1$, the unique solution is $\tilde{\pi}_0 = \frac{\pi_0}{1-\pi_1}$. From **(A)** it follows that $\tilde{\pi}_0 < 1$. Similar reasoning applies to the second identity. □

This lemma motivates estimates of the true Type I and Type II errors. For any classifier f, we may express the contaminated Type I and Type II errors as

$$\begin{aligned}\tilde{R}_0(f) &= \tilde{P}_0(f(X)=1)\\ &= (1-\tilde{\pi}_0)R_0(f) + \tilde{\pi}_0(1-\tilde{R}_1(f)) \qquad (8)\\ \tilde{R}_1(f) &= \tilde{P}_1(f(X)=0)\\ &= (1-\tilde{\pi}_1)R_1(f) + \tilde{\pi}_1(1-\tilde{R}_0(f)), \qquad (9)\end{aligned}$$

where Equations (8) and (9) follow from Lemma 2. By solving for $R_0(f)$ and $R_1(f)$ in (8) and (9), we find

$$R_0(f) = \frac{\tilde{R}_0(f) - \tilde{\pi}_0(1-\tilde{R}_1(f))}{1-\tilde{\pi}_0} = 1 - \tilde{R}_1(f) - \frac{1-\tilde{R}_0(f)-\tilde{R}_1(f)}{1-\tilde{\pi}_0} \qquad (10)$$

$$R_1(f) = \frac{\tilde{R}_1(f) - \tilde{\pi}_1(1-\tilde{R}_0(f))}{1-\tilde{\pi}_1} = 1 - \tilde{R}_0(f) - \frac{1-\tilde{R}_1(f)-\tilde{R}_0(f)}{1-\tilde{\pi}_1}. \qquad (11)$$

We can estimate $\tilde{R}_0(f)$ and $\tilde{R}_1(f)$ from the training data. Therefore, if we can estimate $\tilde{\pi}_0$ and $\tilde{\pi}_1$, then we can estimate $R_0(f)$ and $R_1(f)$, and thereby design a classifier. This approach was analyzed in Scott et al. [41]. In Sec. 7 we describe another approach to classifier design based on surrogate loss minimization that also relies on estimates of $\tilde{\pi}_0$ and $\tilde{\pi}_1$. In the next section we describe a framework that is used to estimate $\tilde{\pi}_0$ and $\tilde{\pi}_1$.

We conclude this section with a converse to Lemma 2:

Lemma 3. *Assume that (6)–(7) hold and $\tilde{P}_1 \neq \tilde{P}_0$. Then $P_1 \neq P_0$ and there exist unique $\pi_1, \pi_0 \in [0,1)$ (namely $\pi_0 = \frac{\tilde{\pi}_0(1-\tilde{\pi}_1)}{1-\tilde{\pi}_1\tilde{\pi}_0}$ and $\pi_1 = \frac{\tilde{\pi}_1(1-\tilde{\pi}_0)}{1-\tilde{\pi}_1\tilde{\pi}_0}$) so that (1)–(2) hold; furthermore,* **(A)** *is satisfied.*

Proof. Assume (6)–(7) hold. Since we assume $\tilde{P}_1 \neq \tilde{P}_0$, it holds that $\tilde{\pi}_1, \tilde{\pi}_0 < 1$. To see that $P_0 \neq P_1$, assume that equality holds. Plugging in (6)–(7) and after straightforward manipulation, we obtain equivalently

$$\frac{1-\tilde{\pi}_1\tilde{\pi}_0}{(1-\tilde{\pi}_1)(1-\tilde{\pi}_0)}\tilde{P}_1 = \frac{1-\tilde{\pi}_1\tilde{\pi}_0}{(1-\tilde{\pi}_1)(1-\tilde{\pi}_0)}\tilde{P}_0,$$

which would contradict the assumption $\tilde{P}_1 \neq \tilde{P}_0$.

Next, in order for identity (1) to hold, by matching distributions in a similar way as in the proof of Lemma 2, we arrive at the equivalent relation $(\tilde{\pi}_0(1-\pi_1) - \pi_0)\tilde{P}_0 = (\tilde{\pi}_0(1-\pi_1) - \pi_0)\tilde{P}_1$. Since $\tilde{P}_1 \neq \tilde{P}_0$, the unique solution is $\pi_0 = \tilde{\pi}_0(1-\pi_1)$. Similarly, for (2) to hold the unique solution is $\pi_1 = \tilde{\pi}_1(1-\pi_0)$. From these we derive the announced expression for π_0, π_1. It is then easy to check that $\pi_0 + \pi_1 - 1 = -\frac{(1-\tilde{\pi}_1)(1-\tilde{\pi}_0)}{1-\tilde{\pi}_1\tilde{\pi}_0} < 0$, so that **(A)** holds. □

Together, Lemmas 2 and 3 imply that for known, distinct uncontaminated distributions $P_0 \neq P_1$, there is an explicit one-to-one correspondence between the contamination proportions (π_1, π_0) of the initial contamination models (1)–(2) under constraint **(A)**, and the proportions $(\tilde{\pi}_1, \tilde{\pi}_0)$ in the representation (6)–(7) (with the only constraint $0 \leq \tilde{\pi}_1, \tilde{\pi}_0 < 1$).

The alternate representations (6)–(7) are *decoupled* in the sense that (6) does not involve P_1, while (7) does not involve P_0. This allows us to estimate $\tilde{\pi}_0$ and $\tilde{\pi}_1$ separately, by reducing to the problem of "mixture proportion estimation" (see next section). It further motivates the mutual irreducibility condition on (P_0, P_1) that, together with **(A)**, ensures that $\tilde{\pi}_0, \tilde{\pi}_1$ are identifiable. The decoupling perspective also allows us to address the following question: Given the contaminated distributions $\tilde{P}_1, \tilde{P}_0$, while (P_0, P_1) are unknown, what are the solutions (π_0, π_1, P_0, P_1) satisfying model (1)–(2)? Obviously, $(0, 0, \tilde{P}_1, \tilde{P}_0)$ is a trivial solution; we will argue that mutual irreducibility ensures that the solution is unique and non-trivial, and furthermore that the resulting P_0, P_1 correspond to maximally denoised versions of $\tilde{P}_1, \tilde{P}_0$. The issues are developed in Section 5. In the next section, we review the work of Blanchard et al. [7].

4. Irreducibility and mixture proportion estimation

Let F, G, and H be distributions on $(\mathcal{X}, \mathfrak{S})$ such that

$$F = (1 - \kappa)G + \kappa H,$$

where $0 \leq \kappa \leq 1$. Mixture proportion estimation is the following problem: given iid realizations from both F and H, estimate κ. This problem was previously addressed by [7], and here we relate the essential definitions and results from that work.

Without additional assumptions, κ is not an identifiable parameter, as noted by Blanchard et al. [7]. In particular, if $F = (1 - \kappa)G + \kappa H$ holds, then any alternate decomposition of the form $F = (1 - \kappa + \delta)G' + (\kappa - \delta)H$, with $G' = (1 - \kappa + \delta)^{-1}((1 - \kappa)G + \delta H)$, and $\delta \in [0, \kappa)$, is also valid. Because we have no direct knowledge of G, we cannot decide which representation is the correct one. Therefore, to make κ identifiable, some additional condition must be assumed. The following definition will be useful.

Definition 4. *Let G, H be probability distributions. We say that G is* irreducible *with respect to H if there exists no decomposition of the form $G = \gamma H + (1 - \gamma)F'$, where F' is some probability distribution and $0 < \gamma \leq 1$. We say that G and H are* mutually irreducible *if G is irreducible with respect to H and vice versa.*

The following was established by Blanchard et al. [7].

Proposition 5. *Let F, H be probability distributions. If $F \neq H$, there is a unique $\kappa^* \in [0, 1)$ and G such that the decomposition $F = (1 - \kappa^*)G + \kappa^* H$ holds, and such that G is irreducible with respect to H. If we additionally define $\kappa^* = 1$ when $F = H$, then in all cases*

$$\kappa^* = \max\{\alpha \in [0, 1] : \exists\, G' \text{ probability distribution: } F = (1 - \alpha)G' + \alpha H\}.$$

By this result, the following is well-defined.

Definition 6. *For any two probability distributions F, H, define*

$$\kappa^*(F|H) := \max\{\alpha \in [0,1] : \exists\, G' \text{ probability distribution: } F = (1-\alpha)G' + \alpha H\},$$

the maximal proportion of H in F.

Clearly, G is irreducible with respect to H if and only if $\kappa^*(G|H) = 0$. It is also interesting to note that $1 - \kappa^*(F|H)$ is an example of a statistical distance. That is, $1 - \kappa^*(F|H)$ is always nonnegative, and is equal to zero if and only if $F = H$, by Proposition 5. Furthermore, Proposition 8 below states that this distance can be expressed in terms of the likelihood ratio, like Kullback-Liebler and other information divergences. This statistical distance has been studied previously for discrete distributions in the analysis of Markov chains [2], where it is called the "separation distance." In general, $\kappa^*(F|H) \neq \kappa^*(H|F)$, so that this is not actually a metric on distributions.

To consolidate the above notions, we state the following corollary which expresses that irreducibility of G with respect to H is sufficient for the mixture proportion to be identifiable.

Corollary 7. *If $F = (1-\gamma)G + \gamma H$, and G is irreducible with respect to H, then $\gamma = \kappa^*(F|H)$.*

Some intuition for κ^* and irreducibility come from the following result. Part of the result is in terms of the receiver operating characteristic (ROC) for the problem of testing the null hypothesis $X \sim H$ against the alternative $X \sim F$. Given a measurable set $S \in \mathfrak{S}$, we can think of S as a rejection region (where the null hypothesis is rejected). Then $H(S)$ is the false positive rate and $F(S)$ is the true positive rate, and the optimal ROC is defined as

$$\beta(\tau) := \sup\{F(S) \,|\, H(S) \leq \tau, S \in \mathfrak{S}\}.$$

The ensuing result follows from Theorem 6 of Blanchard et al. [7].

Proposition 8 (Blanchard et al. [7])**.**

$$\kappa^*(F|H) = \inf_{S \in \mathfrak{S}, H(S) > 0} \frac{F(S)}{H(S)} = \inf_{\tau \in [0,1)} \left\{ \frac{1-\beta(\tau)}{1-\tau} \right\}.$$

If f and h are densities of F and H, respectively, with respect to a common dominating measure, then

$$\kappa^*(F|H) = \operatorname*{ess\,inf}_{x \in \mathrm{supp}(H)} \frac{f(x)}{h(x)}.$$

Proof. The first two identities are established by Blanchard et al. [7]. See also [39]. The proof of the first identity is very similar to the proof of the third identity given below. Intuition for the second identity comes from the first identity and the observation that the optimal ROC is concave. To prove the third identity, let

$$\gamma^* = \operatorname*{ess\,inf}_{x \in \mathrm{supp}(H)} \frac{f(x)}{h(x)}.$$

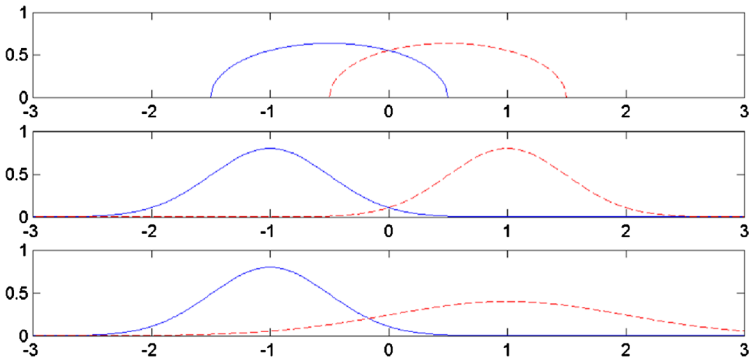

FIG 1. *Three one-dimensional examples that illustrate assumption* **(C)**. *In each example (row),* P_0 *is on the left (solid line) and* P_1 *on the right (dotted line). In the first two examples, mutual irreducibility holds, but in the third example it does not. See text for details.*

We need to show (i) $\exists g$ such that $f = (1-\gamma^*)g + \gamma^* h$, and (ii) if $\gamma > \gamma^*$, then no such g exists. To see (i), take $g = (f - \gamma^* h)/(1-\gamma^*)$, which clearly integrates to one, and is a.s. nonnegative by definition of γ^*. To see (ii), suppose that for some $\gamma > \gamma^*$, there exists a density g with $f = (1-\gamma)g + \gamma h$. Then for all x such that $h(x) > 0$,

$$\frac{f(x)}{h(x)} = \gamma + (1-\gamma)\frac{g(x)}{h(x)} \geq \gamma > \gamma^*,$$

which contradicts the definition of γ^*. □

An alternate proof of the last statement, based on properties of ROC curves of likelihood ratio tests, is given in an appendix.

The result $\kappa^*(F|H) = \inf_{S \in \mathfrak{S}, H(S)>0} \frac{F(S)}{H(S)}$ motivates the universally consistent estimator of κ^* due to [7], reviewed below in Section 6. The second identity, which states that $\kappa^*(F|H)$ is the slope of the optimal ROC at its right end-point, motivates a more practical estimator discussed in Section 11.

Proposition 8 makes it possible to check irreducibility for certain distributions. For example, $\kappa^*(G|H) = 0$ whenever the support of G does not contain the support of H. Irreducibility is also possible even if G and H have the same support, as in the case where G and H are Gaussian distributions with different means, and the variance of H is no more than the variance of G. This follows easily from the density ratio characterization of κ^*.

Proposition 8 also makes it easy to check *mutual* irreducibility for various distributions P_0 and P_1. Indeed, two continuous distributions are mutually irreducible iff the (essential) infimum and supremum of their density ratio are 0 and ∞, respectively. Figure 1 shows three examples where $\mathcal{X} = \mathbb{R}$. In the first example, P_0 and P_1 are such that the support of one is not contained in the support of the other, and therefore mutual irreducibility is satisfied. In the second example, P_0 and P_1 are Gaussian distributions with equal variances and

unequal means. By plugging in the formulas for the Gaussian densities, it is easy to verify that mutual irreducibility again holds. In the third example, P_0 and P_1 are again Gaussian densities with unequal means, but this time with unequal variances. In this case, it is again not hard to show that $\kappa^*(P_0|P_1) = 0$, but $\kappa^*(P_1|P_0) > 0$, where P_1 has the larger variance. Thus, mutual irreducibility does not hold in this case. We do note, however, that $\kappa^*(P_1|P_0)$ tends to zero very fast as the means move apart.

5. Mutual irreducibility: Sufficiency, necessity, and maximal denoising

We argue that mutual irreducibility of P_0 and P_1 is both necessary and sufficient for identifiability of the elements (π_0, π_1, P_0, P_1) of the contamination models, and relate it to the notion of maximal denoising of the contaminated distributions. Since our focus in this section is identifiability and not estimation, our discussion is at the population level.

5.1. Sufficiency of mutual irreducibility for identifiability

Recalling the result of Lemma 2, the distributions $\tilde{P}_0$ and $\tilde{P}_1$ can be written

$$\tilde{P}_0 = (1 - \tilde{\pi}_0)P_0 + \tilde{\pi}_0\tilde{P}_1$$
$$\tilde{P}_1 = (1 - \tilde{\pi}_1)P_1 + \tilde{\pi}_1\tilde{P}_0.$$

By Corollary 7, we can identify $\tilde{\pi}_0$ and $\tilde{\pi}_1$ provided the following condition holds:

(B) P_0 is irreducible with respect to $\tilde{P}_1$ and P_1 is irreducible with respect to $\tilde{P}_0$.

We prefer an irreducibility assumption based on the true class-conditional distributions, and so introduce the following:

(C) P_0 and P_1 are mutually irreducible.

Note that it follows from assumption **(C)** that $P_0 \neq P_1$, which is a hypothesis of Lemma 2. We now establish that **(C)** and **(B)** are essentially equivalent.

Lemma 9. *P_0 is irreducible with respect to $\tilde{P}_1$ if and only if P_0 is irreducible with respect to P_1 and $\pi_1 < 1$. The same statement holds when exchanging the roles of the two classes. In particular, under assumption* **(A)**, **(C)** *is equivalent to* **(B)**.

Proof. This will be a proof by contraposition. Assume first that P_0 is not irreducible with respect to $\tilde{P}_1$. Then there exists a probability distribution Q' and $0 < \gamma \leq 1$ such that

$$P_0 = \gamma\tilde{P}_1 + (1 - \gamma)Q'.$$

Now, plugging in Equation (2) for $\tilde{P}_1$ yields

$$P_0 = \gamma((1-\pi_1)P_1 + \pi_1 P_0) + (1-\gamma)Q'.$$

Solving for P_0 produces

$$P_0 = (1-\beta)Q' + \beta P_1,$$

where $\beta = \gamma(\frac{1-\pi_1}{1-\gamma\pi_1})$. Now, in the case where $\pi_1 < 1$, then $1-\gamma\pi_1 > 0$, and $\gamma - \gamma\pi_1 > 0$. Since $0 < \gamma \leq 1$, we deduce $0 < \beta \leq 1$, so that P_0 is not irreducible with respect to P_1.

Conversely, assume by contradiction that P_0 is not irreducible with respect to P_1, i.e., there exists a decomposition $P_0 = \gamma P_1 + (1-\gamma)Q'$ with $\gamma > 0$. Then the decomposition $P_0 = \beta\tilde{P}_1 + (1-\beta)Q'$ holds with $\beta = \frac{\gamma}{\gamma+(1-\pi_1)(1-\gamma)} \in (0,1]$, so that P_0 is not irreducible with respect to $\tilde{P}_1$. Finally, in the case $\pi_1 = 1$, we have $\tilde{P}_1 = P_0$, in which case, trivially, P_0 is not irreducible with respect to $\tilde{P}_1$ either. □

The following corollary summarizes the discussion of sufficiency.

Corollary 10. *If* **(A)** *and* **(C)** *hold, then* $\pi_0 = \frac{\tilde{\pi}_0(1-\tilde{\pi}_1)}{1-\tilde{\pi}_1\tilde{\pi}_0}$ *and* $\pi_1 = \frac{\tilde{\pi}_1(1-\tilde{\pi}_0)}{1-\tilde{\pi}_1\tilde{\pi}_0}$, *where* $\tilde{\pi}_0 = \kappa^*(\tilde{P}_0|\tilde{P}_1)$ *and* $\tilde{\pi}_1 = \kappa^*(\tilde{P}_1|\tilde{P}_0)$.

Thus, π_0 and π_1 are explicit functions of $\tilde{P}_0$ and $\tilde{P}_1$ under **(A)** and **(C)**. It follows that P_0 and P_1 can then be recovered by solving the identities (6)–(7). In fact, using these identities, it is easy to check that a slightly stronger statement holds: for any arbitrary given $\tilde{P}_0 \neq \tilde{P}_1$, there is a unique solution (π_0, π_1, P_0, P_1) of (1)–(2) satisfying **(A)** and **(C)**. For short, we call this solution the *unique mutually irreducible solution* of the problem (condition **(A)** being tacitly required.) The uniqueness and various properties of this particular solution will be explored in more detail in Theorem 12 below; in the next Section, we first argue that conditions **(A)** and **(C)** are necessary for decontamination in a certain sense.

5.2. Necessity

As noted earlier, given $\tilde{P}_0 \neq \tilde{P}_1$, there are in general many (π_0, π_1, P_0, P_1) solving equations (1)–(2), so that decontamination is not well-defined in the absence of additional conditions. Requesting mutual irreducibility of (P_0, P_1) is one way to ensure unicity of the solution, and also has an interpretation in terms of maximum denoising (see Theorem 12 below). But is it in any way a natural assumption? We now argue that this condition is also the only one ensuring some relatively natural properties of the decontamination operation.

We introduce some additional notation: let $\mathfrak{P}$ denote the set of probability distributions on $\mathcal{X}$. Denote $\mathfrak{P}^2_*$ the set of couples $(P,Q) \in \mathfrak{P}^2$ with $P \neq Q$. We denote ψ the contamination operator from $[0,1]^2 \times \mathfrak{P}^2$ to $\mathfrak{P}^2$, with $\psi(\pi_0, \pi_1, P_0, P_1) = (\tilde{P}_0, \tilde{P}_1)$ given by (1)–(2).

Let ϕ denote a decontamination operator, i.e., a function from a subset of $\mathfrak{P}^2$ to $[0,1]^2 \times \mathfrak{P}^2$ such that $\phi(\tilde{P}_0, \tilde{P}_1)$ returns a solution of (1)–(2), in other words $\psi \circ \phi$ is the identity on the domain of ϕ. We further denote $\phi := (\phi_\pi, \phi_P)$, where $\phi_\pi(\tilde{P}_0, \tilde{P}_1)$ are the solution contamination weights and $\phi_P(\tilde{P}_0, \tilde{P}_1)$ are the solution source distributions. Finally, given a decontamination operator ϕ, call the image of ϕ_P the set of *ϕ-sources* – this is the set of probability distribution couples that are considered as the uncontaminated sources by the operator ϕ in at least one configuration of observed contaminated distributions.

Theorem 11. *Let ϕ denote a decontamination operator satisfying the following conditions:*

(i) Universality: the domain of ϕ is $\mathfrak{P}^2_*$;

(ii) Symmetry: if $\phi(\tilde{P}_0, \tilde{P}_1) = (\pi_0, \pi_1, P_0, P_1)$, then $\phi(\tilde{P}_1, \tilde{P}_0) = (\pi_1, \pi_0, P_1, P_0)$;

(iii) Continuity of recovered contamination weights: for any fixed $P_0 \neq P_1$, the mapping

$$(\pi_0, \pi_1) \in \left\{(\pi_0, \pi_1) \in [0,1]^2; \pi_0 + \pi_1 < 1\right\} \mapsto \phi_\pi(\psi(\pi_0, \pi_1, P_0, P_1))$$

is continuous ;

(iv) Stability of recovered sources: for any ϕ-source (P_0, P_1), there exists $\epsilon > 0$ such that for all $\pi_0, \pi_1 \leq \epsilon$:

$$\phi_P(\psi(\pi_0, \pi_1, P_0, P_1)) = (P_0, P_1)\,. \tag{12}$$

Then $\phi(\tilde{P}_0, \tilde{P}_1)$ must be the unique mutually irreducible solution of (1)–(2) for all $\tilde{P}_0 \neq \tilde{P}_1$.

The interpretation of this result is that mutual irreducibility is a necessary condition for decontamination if conditions (i) to (iv) are required. Condition (i) states that the decontamination operation should be defined on the full domain of possible (distinct) observed distributions and can thus be seen as a universality condition. Condition (ii) is a natural symmetry requirement. Condition (iii) is a continuity assumption (changing the mixing weights by an arbitrarily small amount should not result in a "jump" in the returned estimated contamination proportions) and condition (iv) is a stability condition (a couple (P_0, P_1) identified as a source should still be output as a source by the decontamination operator under small enough mutual mixing proportions.)

Remark. Removing one of the "natural" requirements (i)–(iv) invalidates the conclusion. For example, restricting decontamination to a certain specific model of sources – say Gaussian distributions – could give rise to a non-mutually irreducible decontamination, coherent within that model but forgoing universality (i). If we remove continuity requirement (iii), we can find a decontamination operator that is not mutually irreducible and satisfies the other conditions by "tiling" the solution space: for any (P_0, P_1) mutually irreducible, any (π_0, π_1) such that $\pi_0 + \pi_1 < 1$, $(\tilde{P}_0, \tilde{P}_1) = \psi(\pi_0, \pi_1, P_0, P_1)$, for $\pi_i \in [\frac{k_i}{n}, \frac{k_i+1}{n})$ (n can be

chosen arbitrarily), define the decontamination ϕ as $\phi = (\phi_\pi, \phi_P)$ with

$$\begin{cases} \phi_\pi(\tilde{P}_0, \tilde{P}_1) := \left(\frac{\pi_0 - k_0/n}{1-(k_0+k_1)/n}, \frac{\pi_1 - k_1/n}{1-(k_0+k_1)/n} \right); \\ \phi_P(\tilde{P}_0, \tilde{P}_1) := \psi\left(\frac{k_0}{n}, \frac{k_1}{n}, P_0, P_1 \right). \end{cases}$$

It is easy to check that $\pi_0 + \pi_1 < 1$ implies $\phi_\pi(\tilde{P}_0, \tilde{P}_1) \in [0,1]^2$ and satisfies **(A)**. Then the above ϕ satisfies (i), (ii) and (iv) but is not the mutually irreducible decontamination. Finally, stability condition (iv) is needed in order to prevent "trivial" decontaminations such as $\phi(\tilde{P}_0, \tilde{P}_1) = (0, 0, \tilde{P}_0, \tilde{P}_1)$, which is obviously continuous. Excluding the everywhere trivial decontamination is not enough, as a decontamination could also be trivial on part of the space only.

5.3. Maximal denoising

To conclude this section, we present a result that rounds out the discussion of the initial and modified contamination models, and mutual irreducibility. In particular, we describe all possible solutions (π_0, π_1, P_0, P_1) to our model equations (1)–(2) when $\tilde{P}_0, \tilde{P}_1$ are given and arbitrary, and an equivalent characterization of the unique mutually irreducible solution. It can be seen as an analogue of Proposition 5 for the label noise contamination models.

Theorem 12. *Let $\tilde{P}_1 \neq \tilde{P}_0$ be two given distinct probability distributions. Denote by Λ the feasible set of quadruples (π_0, π_1, P_0, P_1) such that* **(A)** *and equations (1)–(2) are satisfied.*

1. *There is a unique quadruple $(\pi_0^*, \pi_1^*, P_0^*, P_1^*) \in \Lambda$ so that* **(C)** *holds.*
2. *Denoting $\tilde{\pi}_0^* := \kappa^*(\tilde{P}_0|\tilde{P}_1) < 1$ and $\tilde{\pi}_1^* := \kappa^*(\tilde{P}_1|\tilde{P}_0) < 1$, it holds*

$$\pi_0^* = \frac{\tilde{\pi}_0^*(1-\tilde{\pi}_1^*)}{1-\tilde{\pi}_1^*\tilde{\pi}_0^*}, \qquad \pi_1^* = \frac{\tilde{\pi}_1^*(1-\tilde{\pi}_0^*)}{1-\tilde{\pi}_1^*\tilde{\pi}_0^*}. \tag{13}$$

3. *The feasible region R for the proportions (π_0, π_1) (that is, the projection of Λ to its first two coordinates, which is also one-to-one), is the closed quadrilateral defined by the intersection of the positive quadrant of $\mathbb{R}^2$ with the half-planes given by*

$$\pi_0 + \pi_1\tilde{\pi}_0^* \leq \tilde{\pi}_0^*, \qquad \pi_1 + \pi_0\tilde{\pi}_1^* \leq \tilde{\pi}_1^*. \tag{14}$$

4. *The mutually irreducible solution $(\pi_0^*, \pi_1^*, P_0^*, P_1^*)$ is also equivalently characterized as:*
   - *the unique maximizer of $(\pi_0 + \pi_1)$ over Λ;*
   - *the unique extremal point of Λ where both of the constraints in (14) are active;*
   - *the unique maximizer over Λ of $\|P_0 - P_1\|_{TV}$, the total variation distance between the source distributions.*

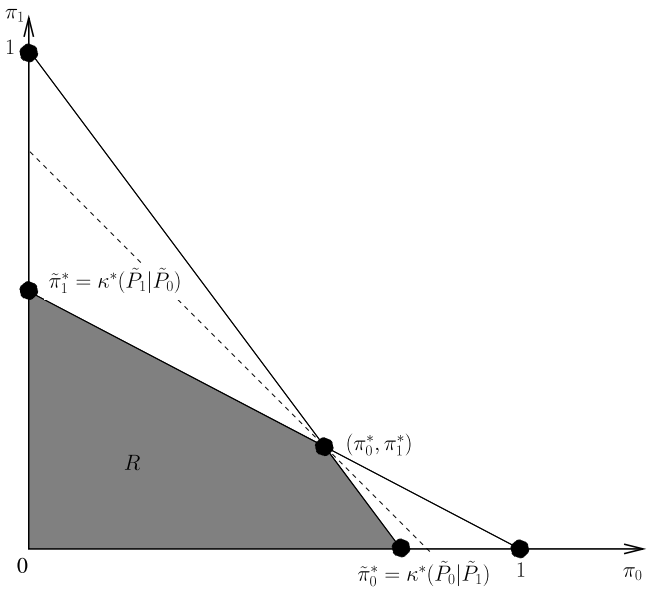

FIG 2. *Geometry of the feasible region* Λ *for proportions* (π_0, π_1) *solutions of the contamination model* (1)–(2)*, when contaminated distributions* $(\tilde{P}_0, \tilde{P}_1)$ *are observed and the true distributions* (P_0, P_1) *are unknown. Each feasible* (π_0, π_1) *corresponds to a single associated solution* (P_0, P_1)*. The extremal point* (π_0^*, π_1^*) *is the unique point corresponding to a mutually irreducible solution* (P_0^*, P_1^*)*. The dashed line indicates the maximal level line* $(\pi_0 + \pi_1) = c$ *intersecting with* Λ*.*

The proof of the theorem relies on the explicit one-to-one correspondence established in Lemmas 2 and 3 between the solutions of the original decomposition (1)–(2) and its decoupled reformulation (6)–(7). The result of Proposition 5 is applied to the decoupled formulation, then pulled back, via the correspondence, in the original representation. The last statement concerning the total variation norm is based on the relation

$$(P_1 - P_0) = (1 - \pi_0 - \pi_1)^{-1}(\tilde{P}_1 - \tilde{P}_0),$$

obtained by subtracting (1) from (2). Therefore, the maximum feasible value of $\|P_1 - P_0\|_{TV}$ corresponds to the maximum of $(\pi_0 + \pi_1)$, i.e., the unique mutually irreducible solution.

The geometrical interpretation of this theorem is visualized on Figure 2. In particular, point 1 of the theorem shows that conditions **(A)** and **(C)** do not restrict the class of possible observable contaminated distributions $(\tilde{P}_1, \tilde{P}_0)$; rather, they ensure in all cases the identifiability of the mixture model. Point 4 indicates that the unique solution satisfying the mutual irreducibility condition **(C)** can be characterized as maximizing the possible total label noise level $(\pi_0 + \pi_1)$, or, still equivalently, the total variation separation of the source probabilities P_0, P_1. In this sense, the mutually irreducible solution can also

be interpreted as *maximal label denoising* or *maximal source separation* of the observed contaminated distributions.

6. Mixture proportion estimation and a rate of convergence

Blanchard et al. [7] present a universally consistent estimator $\widehat{\kappa}$ of $\kappa^*(F|H)$. We review this estimator below. They also establish a "no free lunch" result stating that no estimator of $\kappa^*(F|H)$ can converge at a fixed rate for all F and H. In this section we also introduce distributional assumptions under which the estimator of Blanchard et al. [7] converges at a known rate.

We begin by reviewing the universally consistent estimator of $\kappa^*(F|H)$ introduced by Blanchard et al. [7]. Let F and H be probability measures on a Borel space $(\mathcal{X}, \mathfrak{S})$. Recall from Proposition 8

$$\kappa^*(F|H) = \inf_{S\in\mathfrak{S}, H(S)>0} \frac{F(S)}{H(S)}.$$

The basic idea is to replace F and H by empirical estimates and take the infimum over a union of VC classes. Thus, consider a sequence of VC classes of sets, $(\mathcal{S}_k)_{k\geq 1}$, with respective (finite) VC dimensions $(V_k)_{k\geq 1}$. Define $\epsilon_i(k,\delta_i) := 3\sqrt{\frac{V_k \log(n_i+1) - \log \delta_i/2}{n_i}}$ for $i = 0, 1$. By the VC inequality, for any $i = 0, 1$, $\delta_i \in (0,1)$, $k \geq 1$ and any distribution Q on $\mathcal{X}$, with probability at least $1-\delta_i$ over the draw of an i.i.d. sample of size n_i according to Q, we have

$$\forall S \in \mathcal{S}_k \quad \left|Q(S) - \widehat{Q}(S)\right| \leq \epsilon_i(k,\delta_i)\,, \tag{15}$$

where $\widehat{Q}$ denotes the empirical distribution built on the sample.

In MPE we have training data

$$X_0^1, \ldots, X_0^{n_0} \overset{iid}{\sim} H, \tag{16}$$

$$X_1^1, \ldots, X_1^{n_1} \overset{iid}{\sim} F. \tag{17}$$

For $k \geq 1$, define

$$\widehat{\kappa}(k,\delta_0,\delta_1) := \inf_{S\in\mathcal{S}_k} \frac{\widehat{F}(S) + \epsilon_1(k,\delta_1)}{(\widehat{H}(S) - \epsilon_0(k,\delta_0))_+} \tag{18}$$

where $(\cdot)_+$ is the max of its argument and zero (the ratio is defined to be ∞ if the denominator is zero), and where $\widehat{F}(S)$ and $\widehat{H}(S)$ are the empirical true positive and false positive probabilities associated with the rejection region S. By the VC inequality and Proposition 8, $\widehat{\kappa}(k,\delta_0,\delta_1)$ is an upper bound on $\kappa^*(F|H)$, with probability at least $1-\delta_0-\delta_1$.

Next, define

$$\widehat{\kappa}(\delta_0,\delta_1) := \inf_{k\geq 1} \widehat{\kappa}(k, \delta_0 k^{-2}, \delta_1 k^{-2}).$$

By the union bound, this is also an upper bound on κ^*, with probability at least $1-2(\delta_0+\delta_1)$, since $\sum_k k^{-2} = \pi^2/6 < 2$. To ensure that this upper bound approaches κ^* as $n_0, n_1 \to \infty$, the sequence $(\mathcal{S}_k)_{k=1}^{\infty}$ is assumed to satisfy the following universal approximation property, which we refer to as **(AP1)**: For any $S^* \in \mathfrak{S}$, and any distribution Q,

$$\liminf_{k\to\infty} \inf_{S\in\mathcal{S}_k} Q(S\Delta S^*) = 0\,,$$

where $S\Delta S^* = S\backslash S^* \cup S^*\backslash S$ is the symmetric set difference.

Finally, $\widehat{\kappa}$ is defined as $\widehat{\kappa} = \widehat{\kappa}(\frac{1}{n_0}, \frac{1}{n_1})$. Blanchard et al. [7] show the following, which makes no assumption on the distributions F and H and thus establishes a universally consistent method for MPE.

Theorem 13 (Blanchard et al. [7])**.** *With probability at least* $1-2(\frac{1}{n_0}+\frac{1}{n_1})$, $\widehat{\kappa} \geq \kappa^*(F|H)$. *Furthermore, if* $(\mathcal{S}_k)_{k=1}^{\infty}$ *satisfies* **(AP1)**, *then* $\widehat{\kappa} \xrightarrow{i.p.} \kappa^*(F|H)$ *as* $\min\{n_0, n_1\} \to \infty$.

It should be noted that the statement of the consistency result of Blanchard et al. [7] contains a slight error. We present a correction to the statement of the consistency result in an appendix; see also Scott [39]. The error/correction does not affect the present work.

We now introduce an assumption on F and H that will ensure a certain rate of convergence for $\widehat{\kappa}$ above. This rate will be used in the next section to establish consistency of a discrimination rule. The support of a distribution Q, denoted $\mathrm{supp}(Q)$, is the smallest closed set whose complement has measure zero.

(D) There exists a distribution G and $\gamma \in [0,1]$ such that $\mathrm{supp}(H) \not\subset \mathrm{supp}(G)$ and $F = (1-\gamma)G + \gamma H$.

The assumption $\mathrm{supp}(H) \not\subset \mathrm{supp}(G)$ clearly implies that G is irreducible with respect to H, and therefore γ in **(D)** is equal to $\kappa^*(F|H)$.

In addition, we adopt a modified approximation condition on the sequence $(\mathcal{S}_k)$, referred to as **(AP2)**: For all G, H with $\mathrm{supp}(H) \not\subset \mathrm{supp}(G)$ there exists $k \geq 1$ and $S \in \mathcal{S}_k$ s.t. $G(S) = 0$ and $H(S) > 0$.

Remark. **(AP1)** requires that the sets in $\mathcal{S}_k$ become increasingly complex, so that $V_k \to \infty$. On the other hand, **(AP2)** does not. For example, if $\mathcal{X} = \mathbb{R}^d$ and $\mathfrak{S}$ is the Borel σ-algebra generated by the standard topology on $\mathbb{R}^d$, **(AP2)** is satisfied taking $\mathcal{S}_1$ to be the VC class of all open balls $\{x : \|x-c\| < r\}, c \in \mathbb{R}^d, r > 0$, and $\mathcal{S}_k = \emptyset$ for $k \geq 2$. In this case, we could even simplify the estimator of κ^* to be $\widehat{\kappa}' := \widehat{\kappa}(1, \frac{1}{n_0}, \frac{1}{n_1})$, and the rate of convergence presented below would still hold (the proof requires only minor modifications). However, we elect to work with the definition of $\widehat{\kappa}$ above to emphasize that the rate of convergence applies to the universally consistent estimator.

Theorem 14. *Suppose* $(\mathcal{S}_k)_{k\geq 1}$ *is chosen to satisfy* **(AP2)**. *If* F *and* H *are such that* **(D)** *holds, then there exists a constant* $C > 0$ *such that for* n_0 *and* n_1

sufficiently large, the estimator $\widehat{\kappa}$ *satisfies*

$$\Pr\left(|\widehat{\kappa}-\kappa^*| \geq C\left[\sqrt{\frac{\log n_0}{n_0}}+\sqrt{\frac{\log n_1}{n_1}}\right]\right) \leq \frac{2}{n_0}+\frac{2}{n_1} \tag{19}$$

where $\kappa^* = \kappa^*(F|H)$.

In the next section, assume $\widehat{\kappa}$ is defined in terms of VC classes satisfying **(AP2)**.

7. Consistent classification with unknown label noise proportions

The consistent estimator of κ^* just discussed provides a clear path to the design of a consistent discrimination rule when the label noise proportions are unknown. The estimator of κ^*, together with Corollary 10, can be combined to give consistent estimators of $\tilde{\pi}_0$ and $\tilde{\pi}_1$ under assumptions **(A)** and **(C)**. Plugging in these estimators, along with empirical estimates of $\tilde{R}_0$ and $\tilde{R}_1$, into Eqns. (10) and (11), yields estimates of R_0 and R_1 that can be shown to converge uniformly over a VC class of classifiers to their true values. By allowing the size of the VC class to grow as the sample size(s) grow, empirical risk minimization can be shown to be a consistent discrimination rule with respect to any performance measure defined in terms of R_0 and R_1. This idea utilizes standard ideas in learning theory and is illustrated for the minmax criterion in Scott et al. [41].

One drawback of empirical risk minimization over VC classes is that it is computationally intractable for most VC classes of interest. In the remainder of this section we establish a computationally tractable consistent discrimination rule based on surrogate risk minimization.

7.1. Problem formulation

Let (X, Y) be random on $\mathcal{X} \times \{0, 1\}$ where $\mathcal{X}$ is a Borel space, and let P denote the probability measure governing (X, Y). Let $\mathcal{M}$ denote the set of decision functions, i.e., the set of measurable functions $\mathcal{X} \to \mathbb{R}$. Every $f \in \mathcal{M}$ induces a classifier $x \mapsto u(f(x))$ where $u(t)$ is the unit step function

$$u(t) := \begin{cases} 1, & t > 0 \\ 0, & t \leq 0. \end{cases}$$

For any $f \in \mathcal{M}$, define the *cost-insensitive P-risk* of f

$$R_P(f) := \mathbb{E}_{(X,Y)\sim P}[\mathbf{1}_{\{u(f(X))\neq Y\}}].$$

Define the *cost-insensitive Bayes P-risk* $R_P^* := \inf_{f\in\mathcal{M}} R_P(f)$. It is well known [15] that for any $f \in \mathcal{M}$, the excess P-risk satisfies

$$R_P(f) - R_P^* = 2\mathbb{E}_X[\mathbf{1}_{\{u(f(X))\neq u(\eta(X)-\frac{1}{2})\}}|\eta(X) - \tfrac{1}{2}|], \tag{20}$$

where $\eta(x) := P(Y = 1 \,|\, X = x)$.

Generalizing the above, for any $\alpha \in (0,1)$ we can define the *α-cost-sensitive P-risk* for any $f \in \mathcal{M}$,

$$R_{P,\alpha}(f) := \mathbb{E}_{(X,Y)\sim P}[(1-\alpha)\mathbf{1}_{\{Y=1\}}\mathbf{1}_{\{f(X)\le 0\}} + \alpha\mathbf{1}_{\{Y=0\}}\mathbf{1}_{\{f(X)>0\}}].$$

The corresponding Bayes risk is $R^*_{P,\alpha} := \inf_{f\in\mathcal{M}} R_{P,\alpha}(f)$, and the analogue to (20) is [38]:

$$R_\alpha(f) - R^*_\alpha = \mathbb{E}_X[\mathbf{1}_{\{u(f(X))\neq u(\eta(X)-\alpha)\}}|\eta(X)-\alpha|]\,. \tag{21}$$

Observe (20) corresponds to the case $\alpha = \frac{1}{2}$.

With this background, we turn to the problem of classification with label noise. We assume $(X, Y, \tilde{Y})$ are jointly distributed, where Y is the true but unobserved label, and $\tilde{Y}$ is the observed but noisy label. As in the rest of the paper, we focus on label noise that is independent of the feature vector X, meaning that the conditional distribution of $\tilde{Y}$ given X and Y depends only on Y.

We would like to minimize $R_P(f)$, but we only have access to data from $\tilde{P}$, the joint distribution of $(X, \tilde{Y})$. Natarajan et al. [32] show that minimizing a cost-*sensitive* $\tilde{P}$-risk is equivalent to minimizing the cost-*insensitive* P-risk. We state and prove an equivalent result which has a simpler proof. In this setting, $\pi_i = \Pr(Y = 1-i \,|\, \tilde{Y} = i)$, $i = 0, 1$. We introduce the following assumption on the amount of label noise, which slightly strengthens **(A)**.

(A') $\pi_0 < \frac{1}{2}$ and $\pi_1 < \frac{1}{2}$.

The following result connects the cost-sensitive $\tilde{P}$-risk to the cost-insensitive P-risk.

Lemma 15. *If* **(A')** *holds, then for any $f \in \mathcal{M}$,*

$$R_P(f) - R^*_P = 2(1-\pi_1-\pi_0)(R_{\tilde{P},\alpha}(f) - R^*_{\tilde{P},\alpha})\,, \tag{22}$$

where $\alpha = (\frac{1}{2} - \pi_0)/(1-\pi_1-\pi_0)$.

Proof. Note that **(A')** ensures $\alpha \in (0,1)$. Define $\tilde{\eta}(x)$ in analogy to $\eta(x)$ by $\tilde{\eta}(x) := \Pr(\tilde{Y} = 1|X = x)$, leading to

$$\begin{aligned}
\eta(x) &= \Pr(Y=1,\tilde{Y}=1|X=x) + \Pr(Y=1,\tilde{Y}=0|X=x)\\
&= \Pr(Y=1|\tilde{Y}=1,X=x)\tilde{\eta}(x) + \Pr(Y=1|\tilde{Y}=0,X=x)(1-\tilde{\eta}(x))\\
&= (1-\pi_1)\tilde{\eta}(x) + \pi_0(1-\tilde{\eta}(x))\\
&= (1-\pi_0-\pi_1)\tilde{\eta}(x) + \pi_0.
\end{aligned}$$

Observe that

$$\begin{aligned}
\eta(x) - \tfrac{1}{2} &= (1-\pi_0-\pi_1)\tilde{\eta}(x) + \pi_0 - \tfrac{1}{2}\\
&= (1-\pi_0-\pi_1)[\tilde{\eta}(x) - \alpha].
\end{aligned}$$

The result follows now from (20) and (21):

$$\begin{aligned} R_P(f) - R_P^* &= 2\mathbb{E}_X\left[\mathbf{1}_{\{u(f(X))\neq u(\eta(x)-\frac{1}{2})\}}|\eta(x)-\tfrac{1}{2}|\right] \\ &= 2(1-\pi_0-\pi_1)\mathbb{E}_X\left[\mathbf{1}_{\{u(f(X))\neq u(\tilde{\eta}(x)-\alpha)\}}|\tilde{\eta}(x)-\alpha|\right] \\ &= 2(1-\pi_1-\pi_0)(R_{\tilde{P},\alpha}(f) - R^*_{\tilde{P},\alpha}). \end{aligned}$$

□

The problem we will address is the construction of a discrimination rule $\widehat{f}_n$ that is computationally tractable, does not know α, π_0, or π_1, and is such that $R_P(\widehat{f}_n) - R_P^* \to 0$ in probability. To achieve this, we develop an algorithm $\widehat{f}_n$ based on surrogate risk minimization such that $R_{\tilde{P},\alpha}(\widehat{f}_n) - R^*_{\tilde{P},\alpha} \to 0$ in probability.

7.2. Surrogate losses

A *loss* is any measurable function $L : \{0,1\} \times \mathbb{R} \to [0,\infty)$. For example, the P-risk is defined in terms of the $0-1$ loss, $L(y,t) = \mathbf{1}_{\{y\neq u(t)\}}$. Given a loss L we define the risk

$$R_{P,L}(f) = \mathbb{E}_{(X,Y)\sim P}[L(Y, f(X))],$$

and the corresponding optimal risk $R^*_{P,L} = \inf_{f\in\mathcal{M}} R_{P,L}(f)$.

A *surrogate loss* is one that is used as a surrogate for another, such as a loss L that is convex in its second argument in lieu of the 0-1 loss. Surrogate losses are common in machine learning because they can often be optimized efficiently, unlike the 0-1 loss and its cost-sensitive variants. The notion of classification calibration was developed to theoretically justify the use of surrogate losses. A loss L is said to be *α-classification calibrated* iff there exists an increasing and continuous function θ with $\theta(0) = 0$ such that for all $f \in \mathcal{M}$,

$$R_{P,\alpha}(f) - R^*_{P,\alpha} \leq \theta(R_{P,L}(f) - R^*_{P,L}).$$

An equivalent and more technical characterization of α-CC is provided by [38], but the above definition suffices for our purposes. The point is that driving the surrogate excess risk to zero drives the target excess risk to zero for α-CC losses, and the former can be accomplished by computationally tractable methods like support vector machines, as shown below.

Any loss L can be expressed as $L(y,t) = \mathbf{1}_{\{y=1\}}L_1(t) + \mathbf{1}_{\{y=0\}}L_0(t)$. Given a loss L and $\alpha \in (0,1)$, define

$$L_\alpha(y,t) := (1-\alpha)\mathbf{1}_{\{y=1\}}L_1(t) + \alpha\mathbf{1}_{\{y=0\}}L_0(t). \tag{23}$$

[38] establishes that L is $\frac{1}{2}$-CC iff L_α is α-CC. Several examples of $\frac{1}{2}$-CC losses are known, so these readily translate to examples of α-CC losses via Eqn. (23). In particular, Bartlett et al. [6] establish that if $L(y,t) = \phi((2y-1)t)$ where ϕ is convex and differentiable at 0 with $\phi'(0) < 0$, then L is $\frac{1}{2}$-CC. This justifies several common losses including the hinge loss ($\phi(z) = \max\{0, 1-z\}$) and the logistic loss ($\phi(z) = \log(1+\exp(-z))$). Combining these ideas with Lemma 15 leads to the following result.

Corollary 16. *Suppose L is $\frac{1}{2}$-CC, assume* **(B)** *is satisfied and let $\alpha = (\frac{1}{2} - \pi_0)/(1 - \pi_1 - \pi_0)$. Then there exists an increasing and continuous function θ with $\theta(0) = 0$ such that for all $f \in \mathcal{M}$,*

$$R_P(f) - R_P^* \leq \theta(R_{\tilde{P},L_\alpha}(f) - R_{\tilde{P},L_\alpha}^*).$$

Natarajan et al. [32] consider the setting where π_0 and π_1 are known. Using the above result, they apply Rademacher complexity analysis to establish performance guarantees for a classification strategy based on regularized empirical risk minimization with a surrogate loss L_α.

7.3. Estimating α

When π_0 and π_1 are unknown, a natural strategy is to base a learning algorithm on a surrogate loss $L_{\widehat{\alpha}}$, where $\widehat{\alpha}$ is an estimate of α. We propose an estimate of the form

$$\widehat{\alpha} = \frac{\frac{1}{2} - \widehat{\pi}_0}{1 - \widehat{\pi}_0 - \widehat{\pi}_1},$$

where $\widehat{\pi}_0$ and $\widehat{\pi}_1$ are estimates based on our previously developed results. In particular, suppose we observe noisy data

$$(X_1, \tilde{Y}_1), \ldots, (X_n, \tilde{Y}_n) \overset{iid}{\sim} \tilde{P},$$

One difference to note going forward is that the sample sizes n_0 and n_1 are now random, whereas before they were considered to be nonrandom. This turns out to be a minor difference; see the proof of Proposition 17 below.

Now, let $\widehat{\tilde{\pi}}_0$ and $\widehat{\tilde{\pi}}_1$ be estimates of $\tilde{\pi}_0$ and $\tilde{\pi}_1$ obtained by applying the estimator $\widehat{\kappa}$ of Section 6 twice. The formulas from Lemma 3 lead to the estimates

$$\widehat{\pi}_0 = \frac{\widehat{\tilde{\pi}}_0(1 - \widehat{\tilde{\pi}}_1)}{1 - \widehat{\tilde{\pi}}_0\widehat{\tilde{\pi}}_1} \quad \text{and} \quad \widehat{\pi}_1 = \frac{\widehat{\tilde{\pi}}_1(1 - \widehat{\tilde{\pi}}_0)}{1 - \widehat{\tilde{\pi}}_0\widehat{\tilde{\pi}}_1}. \tag{24}$$

By Corollary 10, if **(A)** and **(C)** hold, then $\tilde{\pi}_0 = \kappa^*(\tilde{P}_0|\tilde{P}_1)$ and $\tilde{\pi}_1 = \kappa^*(\tilde{P}_1|\tilde{P}_0)$, and consequently $\widehat{\pi}_0$ and $\widehat{\pi}_1$ are consistent estimators of π_0 and π_1, respectively. For some of our subsequent analysis, we actually want $\widehat{\pi}_0$ and $\widehat{\pi}_1$ (and therefore $\widehat{\alpha}$) to converge at a known rate. Hence, we want $\tilde{P}_0$ and $\tilde{P}_1$ to satisfy assumption **(D)** in both directions. The following assumption, which strengthens **(C)**, is sufficient for this purpose.

(C') $\text{supp}(P_0) \not\subset \text{supp}(P_1)$ and $\text{supp}(P_1) \not\subset \text{supp}(P_0)$.

This assumption is reasonable in many classification problems. It essentially says that for each of the two (noise-free) classes, there exist patterns belonging to that class that could not possibly be confused with patterns from the other class. We have the following.

Proposition 17. *If* **(A')** *and* **(C')** *hold, then there exist* $C_1, C_2 > 0$ *such that for* n *sufficiently large,*

$$\Pr\left(|\widehat{\alpha} - \alpha| \geq C_1\sqrt{\frac{\log n}{n}}\right) \leq \frac{C_2}{n}.$$

Proof. **(A')** implies $\pi_0 + \pi_1 < 1$, and by **(C')**, P_0 and P_1 are mutually irreducible. Thus Corollary 10 implies $\tilde{\pi}_0 = \kappa^*(\tilde{P}_0|\tilde{P}_1)$ and $\tilde{\pi}_1 = \kappa^*(\tilde{P}_1|\tilde{P}_0)$. We will apply Theorem 14 to both of the estimators $\widehat{\tilde{\pi}}_0$ and $\widehat{\tilde{\pi}}_1$. To verify the assumptions of that theorem, we need to verify **(D)** for both $(F, H) = (\tilde{P}_1, \tilde{P}_0)$ and $(F, H) = (\tilde{P}_0, \tilde{P}_1)$. We will show **(D)** for $(F, H) = (\tilde{P}_1, \tilde{P}_0)$, the other case being similar. From (7), it suffices to show $\mathrm{supp}(\tilde{P}_1) \not\subset \mathrm{supp}(P_0)$. But this holds because $\tilde{P}_1 = (1 - \pi_1)P_1 + \pi_1 P_0$ (see Eqn. (2)) and $\mathrm{supp}(P_1) \not\subset \mathrm{supp}(P_0)$ and $\pi_1 < 1$. We can now apply Theorem 14 to both $\widehat{\tilde{\pi}}_0$ and $\widehat{\tilde{\pi}}_1$. To do so, since n_0 and n_1 are nonrandom in that result, we must condition on n_0 and n_1, and appeal to the fact that, with high probability, n_0 and n_1 are proportional to n. In particular, if $\tilde{q} = \Pr(\tilde{Y} = 1)$, then the relative Chernoff bound implies that with high probability, $n_1 \in (\frac{1}{2}\tilde{q}n, \frac{3}{2}\tilde{q}n)$ and $n_0 \in (\frac{1}{2}(1 - \tilde{q})n, \frac{3}{2}(1 - \tilde{q})n)$. Conditioning on n_0 and n_1 belonging to these intervals, Theorem 14 implies that both $\widehat{\tilde{\pi}}_1$ and $\widehat{\tilde{\pi}}_0$ converge at rates that are $O(\sqrt{\log n/n})$. These rates lead to similar rates for $\widehat{\pi}_1$ and $\widehat{\pi}_0$ (note in particular that assumption **(A')** implies that the denominators in (24) are bounded away from 0 by a fixed margin with large probability for n large enough, independently of π_1, π_0). This in turn leads to the desired rate for $\widehat{\alpha}$. □

7.4. Algorithm

We now introduce a consistent classification procedure based on surrogate losses in the case of unknown label noise proportions. The algorithm relies on the framework of reproducing kernel Hilbert spaces. Thus, let $\mathcal{H}$ be a RKHS, and let L be a loss for binary classification. We say that L is Lipschitz if $L(y, t)$ is a Lipschitz function of t for each y. The algorithm returns the classifier

$$\widehat{f}_n = \arg\min_{f \in \mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} L_{\widehat{\alpha}}(\tilde{Y}_i, f(X_i)) + \lambda_n \|f\|_{\mathcal{H}}^2, \tag{25}$$

where $L_{\widehat{\alpha}}$ is the $\widehat{\alpha}$-weighted cost-sensitive loss associated with L, as defined in (23). For example, if $L(y, t) = \max\{0, 1 - (2y - 1)t\}$ is the hinge loss, $\widehat{f}_n$ is a cost-sensitive support vector machine.

7.5. First consistency result

We will assume that the reproducing kernel k associated with $\mathcal{H}$ is universal and bounded [43]. The former property implies that elements of the RKHS can get arbitrarily close to the Bayes risk. The latter property states that $\sup_x k(x, x) =: B^2 < \infty$. The Gaussian kernel is an example satisfying both of these properties.

Theorem 18. *Assume* **(A')** *and* **(C')** *hold, that the reproducing kernel associated with $\mathcal{H}$ is universal and bounded, and that L is a Lipschitz, $\frac{1}{2}$-CC loss. Let $\lambda_n > 0$ tend to zero as $n \to \infty$ such that $\lambda_n \sqrt{n/\log n} \to \infty$. Then*

$$R_P(\widehat{f}_n) - R_P^* \to 0 \qquad \text{in probability},$$

as $n \to \infty$.

7.6. Alternate consistency result with clippable losses

It is possible to establish a consistency theorem without requiring a rate of convergence on $\widehat{\alpha}$ (thus only requiring the milder condition **(C)** rather than **(C')**), at the expense of treating a more narrow class of losses.

A *T-clippable loss* $L(y,t)$ (see 43, Section 2.2) satisfies the following property:

$$\forall y \in \{0,1\}\ , \forall t \in \mathbb{R}\ \ : L(y, \mathrm{Clip}_T(t)) \leq L(y,t)\,,$$

where $\mathrm{Clip}_T(t) := \min(T, \max(-T,t))$. It is shown by [43], Lemma 2.23, that a convex loss is T-clippable iff $\forall y \in \{0,1\}$, the function $t \in \mathbb{R} \mapsto L(y,t)$ admits a minimum which is attained for some $t \in [-T,T]$. As a consequence, many common surrogate losses are clippable; for instance the hinge loss, the squared loss and the truncated squared loss are 1-clippable. On the other hand, the logistic and the exponential losses are not clippable.

Theorem 19. *Assume* **(A')** *and* **(C)** *hold, that the reproducing kernel associated with $\mathcal{H}$ is universal and bounded, and that L is a Lipschitz, T-clippable, $\frac{1}{2}$-CC loss. Let $\lambda_n > 0$ tend to zero as $n \to \infty$ such that $\lambda_n n \to \infty$. Define $\breve{f}_n := \mathrm{Clip}_T(\widehat{f}_n)$, where $\widehat{f}_n$ is defined by (25). Then*

$$R_P(\breve{f}_n) - R_P^* \to 0 \qquad \text{in probability,}$$

as $n \to \infty$.

8. A more general analysis of co-training

Co-training is a model for binary classification in which the feature vector can be partitioned into two sets of variables, called "views." The critical assumption of co-training is that the views are conditionally independent, given the class label. We refer to this assumption as the *co-training assumption*. Blum and Mitchell [8] show that under this assumption, the optimal classifier can be learned from *unlabeled data* only, provided the learner has access to a "weakly-useful predictor," which is a classifier that, roughly speaking, is at least slightly better than random guessing. The basic idea is to apply the weakly-useful predictor to one of the views to generate noisy labels for the other view. By the co-training assumption, the problem now reduces to classification with label noise. The original analysis assumes that the true label is a deterministic function of either view. Our framework allows us to relax this assumption.

To state our result, we assume that the feature vector X and label Y are jointly distributed with joint distribution Q. Let P_0 and P_1 be the class conditional distributions of Q. Furthermore, let X be expressed as (X^A, X^B), representing the two views. Under the co-training assumption, X^A and X^B are conditionally independent given Y. The unlabeled training data are $X_1, \ldots, X_n$. A *weakly-useful classifier* is a classifier h such that $0 < Q(\{x : h(x) = 1\}) < 1$ and $q_0(h) + q_1(h) < 1$, where

$$q_i(h) = Q(Y = 1 - i \,|\, h(X) = i).$$

Theorem 20. *Let h^A be a known weakly-useful classifier based on view A. Assume that the class-conditional distributions of X^B are mutually irreducible, and let $X_1, \ldots, X_n$ be iid. Under the co-training assumption, there exists a classification algorithm $\widehat{f}_n$ such that $R_Q(\widehat{f}_n) \to R^*_Q$ in probability as $n \to \infty$.*

Proof. Consider the data set

$$(X_1^B, \tilde{Y}_1), \ldots, (X_n^B, \tilde{Y}_n),$$

where $\tilde{Y}_i = h^A(X^A)$. By the co-training assumption, the class-conditional distribution of $\tilde{Y}$ given X^B and the true label Y is not dependent on X^B. Therefore we have the setting of a label noise problem. Since $0 < Q(\{x^A : h^A(x^A) = 1\}) < 1$, the numbers of examples n_0 and n_1 with each noisy label grow with n. Furthermore, the contamination probabilities

$$\pi_i = \Pr(Y = 1 - i \,|\, \tilde{Y} = i)$$

are just $\pi_i = q_i(h^A)$. Since h^A is weakly-useful, we have that $\pi_0 + \pi_1 < 1$. We also have mutual irreducibility for this label noise problem, by assumption. Therefore, a consistent classification rule exists by the construction in Scott et al. [41]. □

The key point is that this result weakens the assumption of deterministic class labels to a mutual irreducibility assumption. The existence of a weakly-useful classifier could be guaranteed, for example, if a small amount of labeled training data was available.

The previous argument relies on the consistent classification rule from [41]. The consistency result for classifiers based on clippable surrogate losses, from earlier in this paper, could also be employed provided the definition of a weakly-useful classifier is strengthened to require that $q_i(h) < \frac{1}{2}$ for each i.

9. Mutual irreducibility and class probability estimation

In this section, we relate mutual irreducibility of P_0 and P_1 to the problem of class probability estimation. Let p_0 and p_1 be densities of P_0 and P_1 with respect to a common dominating measure. Further assume that the feature

vector X and label Y are jointly distributed with joint distribution P, and that $q := P(Y = 1) \in (0, 1)$. The posterior probability that $Y = 1$ is denoted

$$\eta(x) := P(Y = 1 \,|\, X = x).$$

The problem of estimating η from data is known as class probability estimation [12, 35]. The most well-known approach to class probability estimation is logistic regression, which posits the model

$$\widehat{\eta}(x) = \frac{1}{1 + \exp\{-(w^T x + b)\}},$$

where w and x have the same dimension, and $b \in \mathbb{R}$. The parameters w and b are fit to the data by maximum likelihood. More generally, estimates for η commonly have the form

$$\widehat{\eta}(x) = \psi^{-1}(h(x))$$

where $\psi : [0, 1] \mapsto \mathbb{R}$ is a *link* function, and h is a decision function of some sort.

Now define

$$\eta_{\min} := \operatorname*{ess\,inf}_{x \in \mathcal{X}} \ \eta(x) \qquad \text{and} \qquad \eta_{\max} := \operatorname*{ess\,sup}_{x \in \mathcal{X}} \ \eta(x).$$

The following result connects the posterior class probability to mutual irreducibility.

Proposition 21. *With the notation defined above,*

$$\eta_{\max} = \frac{1}{1 + \frac{1-q}{q} \kappa^*(P_1 | P_0)} \tag{26}$$

and

$$\eta_{\min} = 1 - \frac{1}{1 + \frac{q}{1-q} \kappa^*(P_0 | P_1)}. \tag{27}$$

Therefore, P_0 and P_1 are mutually irreducible if and only if $\eta_{\min} = 0$ and $\eta_{\max} = 1$.

Proof. By Bayes' rule, it is true that almost everywhere,

$$\begin{aligned} \eta(x) &= \frac{q p_1(x)}{q p_1(x) + (1 - q) p_0(x)} \\ &= \frac{1}{1 + \frac{1-q}{q} \frac{p_0(x)}{p_1(x)}}. \end{aligned}$$

Equation (26) now follows from Proposition 8. Similarly, we have (almost everywhere)

$$\begin{aligned} \eta(x) &= 1 - \frac{(1 - q) p_0(x)}{(1 - q) p_0(x) + q p_1(x)} \\ &= 1 - \frac{1}{1 + \frac{q}{1-q} \frac{p_1(x)}{p_0(x)}}. \end{aligned}$$

Now (27) follows from Proposition 8. The final statement follows from (26) and (27) and the definition of mutual irreducibility. □

Thus, estimates of $\kappa^*(P_0|P_1)$ and $\kappa^*(P_1|P_0)$ could be used to inform choices about the design of the link function (e.g., its domain) and model class of decision functions.

Proposition 21 also suggest another possible approach to mixture proportion estimation. Suppose $\widehat{\eta}$ is an estimator for η that is consistent with respect to the supremum norm, and let $\widehat{q}$ be the empirical estimate of q based on a random sample from P. Inverting Equation (26),

$$\widehat{\kappa}_{1,0} := \left(\frac{1}{\sup_{x\in\mathcal{X}} \widehat{\eta}(x)} - 1\right) \frac{\widehat{q}}{1-\widehat{q}},$$

is a consistent estimate of $\kappa^*(P_1|P_0)$. Similar remarks apply to $\kappa^*(P_0|P_1)$. Although this suggests that class probability estimation solves mixture proportion estimation in the binary classification context, we note that sup-norm consistency will require distributional assumptions, and therefore the distribution-free estimator of Blanchard et al. [7] is a more general solution.

All of the above observations were present in our original technical report on this topic [42]. Since then, Liu and Tao [25] and Menon et al. [30] have further explored the idea of estimating label noise proportions from the minimum and maximum of the contaminated class probability function. In particular, we note the following.

An immediate corollary of Proposition 21 is the following. Let $\tilde{P}$ be the joint distribution on $(X, \tilde{Y})$, $\tilde{q} = \tilde{P}(\tilde{Y} = 1)$, $\tilde{\eta}(x) = \tilde{P}(Y = 1 \,|\, X = x)$, and let $\tilde{\eta}_{\max}$ and $\tilde{\eta}_{\min}$ denote the essential supremum and infimum of $\tilde{\eta}$. Further let $\tilde{P}_1$ and $\tilde{P}_0$ denote the class conditional distributions of X given $\tilde{Y} = 1, 0$, respectively.

Corollary 22. *Consider the setting of the previous paragraph. If* **(A)** *and* **(C)** *hold, then*

$$\pi_0 = \frac{\tilde{\eta}_{\min}(\tilde{\eta}_{\max} - \tilde{q})}{\tilde{q}(\tilde{\eta}_{\max} - \tilde{\eta}_{\min})} \tag{28}$$

and

$$\pi_1 = \frac{(1 - \tilde{\eta}_{\max})(\tilde{q} - \tilde{\eta}_{\min})}{(1 - \tilde{q})(\tilde{\eta}_{\max} - \tilde{\eta}_{\min})}. \tag{29}$$

Proof. By Proposition 21, we have that

$$\tilde{\eta}_{\max} = \frac{1}{1 + \frac{1-\tilde{q}}{\tilde{q}}\kappa^*(\tilde{P}_1|\tilde{P}_0)}$$

and

$$\tilde{\eta}_{\min} = 1 - \frac{1}{1 + \frac{\tilde{q}}{1-\tilde{q}}\kappa^*(\tilde{P}_0|\tilde{P}_1)}.$$

The result now follows from these equations, Corollary 10, and algebra. □

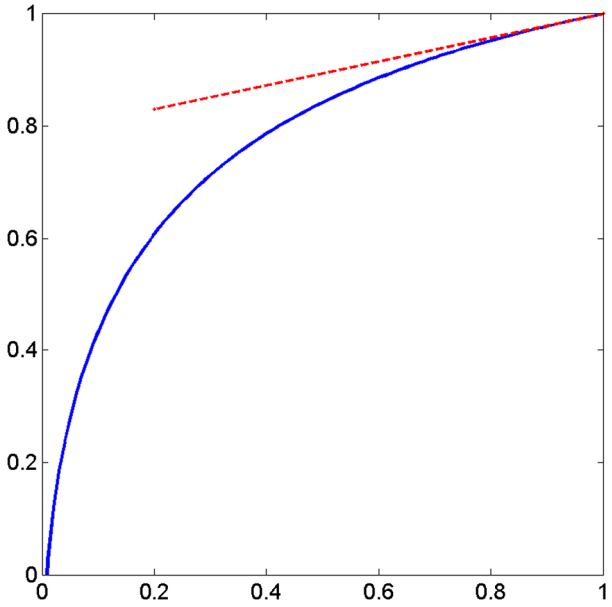

FIG 3. *$\kappa^*(F|H)$ is the slope of the optimal receiver operating characteristic for testing $H_0 : X \sim H$ versus $H_1 : X \sim F$ at its right endpoint.*

Note that $\tilde{q}$ is easily estimated from the fraction of contaminated training examples with $\tilde{Y} = 1$. Therefore, estimates of $\tilde{\eta}_{\max}$ and $\tilde{\eta}_{\min}$ lead directly to estimates of the contamination proportions π_0 and π_1. This approach to estimating label noise proportions is explored experimentally below, where it is compared with the ROC-based estimator.

Menon et al. [30] adopt the conditions $\eta_{\max} = 1$ and $\eta_{\min} = 0$ together with **(A)** as their identifiability conditions for label noise under the contamination model. From the above discussion, these conditions are clearly equivalent to ours. Liu and Tao [25] consider the label flipping model for label noise. They consider an equivalent sufficient condition based on η in that context. Connections with our mutual irreducibility assumption are noted in each of these works.

10. Implementation of estimators

The ROC characterization from Proposition 8 says that κ^* is the minimum slope of any line passing through the point $(1, 1)$ in ROC space and any other point on the optimal ROC. If the optimal ROC happens to be concave, this is the slope of the ROC at its right end-point. See Fig. 3.

Motivated by this idea, we suggest the following practical algorithm for MPE. First, split each of the two samples (1) and (2) into two portions according to a common ratio. Using the first portion of each data set, run a universally consistent classification algorithm that yields a full ROC. In our implementation, we

run kernel logistic regression (KLR) with a Gaussian kernel, and vary the threshold on the posterior probability estimate to obtain an ROC. Note that KLR is run on the contaminated data. The bandwidth and regularization parameters of KLR are set using cross-validation.

Using the second half of each sample, construct conservative estimates (as in Eqn. (18)) of the ROC for a discrete set of thresholds on the KLR posterior probability function. To obtain these conservative estimates, we do not use the empirical error plus or minus a VC bound. Instead, we use direct binomial tail inversion (also known as one-sided exact Clopper-Pearson confidence interval), which is the tightest possible deviation bound for a binomial random variable [22]. Using these conservative estimates, we then compute the minimum slope of all line segments joining points on the ROC to the point $(1, 1)$.

We also considered an alternative approach to estimating the label noise proportions, based on class probability estimation as discussed in Section 9. As in the preceding estimator, we split each sample into two portions, and used the first portion of each sample to train a KLR estimate of the class probability function $\tilde{\eta}$. We then used the second portion of each sample to estimate the minimum and maximum values of $\tilde{\eta}$, which we then plugged into the formulas (28)–(29) to obtain estimates of π_0 and π_1. To obtain some robustness to outliers, we estimated the maximum and minimum using the 99th and 1st percentiles, respectively, as suggested by [30].

The estimates based on the ROC method are denoted $\widehat{\pi}_0^{\mathrm{roc}}$ and $\widehat{\pi}_1^{\mathrm{roc}}$, while the estimates based on class probability estimation are denoted $\widehat{\pi}_0^{\mathrm{cpe}}$ and $\widehat{\pi}_1^{\mathrm{cpe}}$. The former estimates are based on a 20/80 split of each sample, and the latter on a 80/20 split, as these seemed to give the best results. The latter ratio was also employed by Menon et al. [30]. A detailed Matlab implementation, reproducing our results, is available at `http://web.eecs.umich.edu/~cscott`.

11. Experiments

To study the performance of the above estimators, we examined the problem of classification with label noise using three data sets. The waveform data set is available from the UCI Repository, and consists of three classes of synthetically generated waveforms. The classes are overlapping, as the Bayes risk for this data set is known to be around 10 %. We generated data for a binary classification problem (using only two of the classes) with label noise proportions π_0 and π_1 specified as in Table 1. Sample sizes of $n_0 = n_1 = 1000$ were chosen. We also used the MNIST handwritten digits data set, digits 3 and 8, with a similar setup as to the waveform data. In this case the sample sizes were $n_0 = n_1 = 2000$.

A third data set comes from nuclear particle classification, where the training data are realistically described by the label noise model. The data are obtained from organic scintillation detectors, which detect both gamma-rays and neutrons, and associate every detected particle with a digitally sampled pulse-shaped waveform [1]. The goal is to classify gamma-ray pulses (class 0) from neutron pulses (class 1). See discussion in Section 1.1. Training data were obtained by measuring particles emitted from a Cf-252 source, which undergoes

TABLE 1
Results for mixture proportion estimation as applied to classification with label noise.

data set	π_0	π_1	$\widehat{\pi}_0^{\text{roc}}$	$\widehat{\pi}_1^{\text{roc}}$	$\widehat{\pi}_0^{\text{cpe}}$	$\widehat{\pi}_1^{\text{cpe}}$
waveform	0.1	0.25	0.0979	0.2792	0.1919	0.1393
waveform	0.15	0.05	0.1437	0.0589	0.0369	0.0831
digits	0.1	0.25	0.1325	0.2573	0.1065	0.0555
digits	0.15	0.05	0.1633	0.0597	0.0191	0.0479
nuclear	N/A	N/A	0.0100	0.0641	0.0151	0.0007

spontaneous decay and emits both neutrons and gamma rays. Data were preprocessed by aligning pulse peaks and by eliminating signals with multiple peaks (corresponding to multiple detected events within a single observation window). Through a special experimental configuration [3], the time of flight (TOF) for each particle hitting the detector was also measured. Since neutrons travel more slowly than gamma-rays, this gives noisy labels by looking only at those particles with TOF in a certain window. Gamma-rays travel at the speed of light, so a data set with mostly gamma-ray pulses was obtained by focusing on those particles with TOFs around the speed of light (TOF $<$ 5 ns). However, neutrons can still have TOFs in this window because they were generated from either a background event or from another fission event that occurred just an instant before the one being measured. A neutron TOF-window was also selected ($45 <$ TOF $<$ 55 ns), and as with the other window, this one will also contain some proportion of gamma-ray pulses. We obtained samples of size $n_0 = n_1 = 3000$ from each window. It is important to keep in mind that in this application, the ground truth π_0 and π_1 are unknown, and it can only be assessed whether our estimates of these quantities are reasonable based on physics knowledge.

The results are reported in Table 1. Regarding the ROC method, the results indicate that this method provides reasonably accurate estimates of the label noise proportions in the four experimental settings where the true proportions are known. These results also suggest that mutual irreducibility can be a reasonable assumption in practice. In the nuclear particle classification problem, although ground truth labels are unavailable, the proportions estimated by the ROC method are at least consistent with the expectation that noisy labels should be relatively rare (given the high rate of Cf-252 fission events relative to the expected rate of background events), and also with the knowledge that neutrons are rarer background events than gamma-rays (i.e., $\pi_0 < \pi_1$).

With regards to the CPE method, the results indicate that the method is sometime accurate, but other times incurs considerable error. We also note that for the nuclear data, π_1 is estimated to be smaller than π_0, which is inconsistent with the knowledge that contaminating neutrons are more rare than contaminating gamma-rays. To further investigate this issue, we formed Table 2. The first two columns are the same as in the previous table, restricted to the waveform and digits data for which ground truth is known. The third and fourth columns show the empirical percentiles of the ground truth values of $\tilde{\eta}_{\min}$ and $\tilde{\eta}_{\max}$, which should ideally be near 0 and 1.

TABLE 2
Percentiles of the true $\tilde{\eta}_{\min}$ and $\tilde{\eta}_{\max}$ for those experiments with ground truth. The 3rd and 4th columns should ideally be zero and 1.

data set	π_0	π_1	$\tilde{\eta}_{\min}$ %ile	$\tilde{\eta}_{\max}$ %ile
waveform	0.1	0.25	0	0.79
waveform	0.15	0.05	0.28	0.99
digits	0.1	0.25	0.07	0.70
digits	0.15	0.05	0.31	0.93

We see that our implementation of the CPE estimator can be both conservative (estimating more noise than is actually present), and overly optimistic (estimating less noise than is present). Indeed, percentiles far from 0 or 1 reflect over optimism. On the other hand, percentiles of exactly 0 and 1 (of which there is one instance in Table 2) are quite likely signs of conservitism. In this case, the empirical values of $\tilde{\eta}$ do not cover the full range $[\tilde{\eta}_{\min}, \tilde{\eta}_{\max}]$.

CPE-based estimators were also studied by Liu and Tao [25], Menon et al. [30], who report more favorable results for this method. We use KLR to estimate the class probabilities, whereas they employ different techniques. Given this discrepancy in findings, the issue warrants further investigation. There are two factors that may favor the ROC-method. First, the ROC method employs uncertainty quantification (on the deviation between true and empirical probabilities) in the form of direct binomial tail inversion when estimating the slope of the ROC at its right endpoint. Similar uncertainty quantification would likely benefit the CPE method and make it less overly optimistic. Second, the ROC method leverages the shape constraint that it is typically concave.

To illustrate the importance of accounting for label noise, we further examine nuclear particle classification. As noted in Section 2, training a classifier on contaminated training data generates the same ROC as training with uncontaminated data, and the real impact of accounting for label noise occurs in performance evaluation. In Fig. 4, the solid curve plots the ROC for the nuclear particle data, using contaminated test data to estimate the false positive and true positive rates. The dotted curve relies on Eqns. (10)–(11) to correct these probabilities, revealing that the classifier actually classifies the particles much more accurately than one would expect if label noise was not accounted for. This makes intuitive sense, because many of the particles from the contaminated test data that appear to be incorrectly classified are actually correctly classified, and just have erroneous labels.

12. Conclusion

We argue that consistent classification with label noise is possible if a majority of the labels are correct on average, and the class-conditional distributions P_0 and P_1 are mutually irreducible. Under these conditions, we leverage results of [7] on mixture proportion estimation to design consistent estimators of the noise proportions. These estimators are applied to establish a consistent discrimination rule based on surrogate loss minimization, although other performance

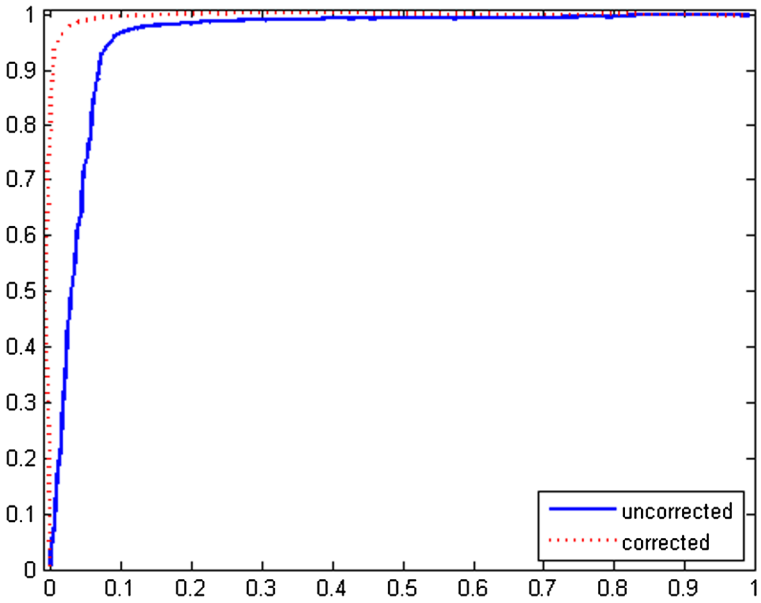

FIG 4. *The ROC for the nuclear classification problem, where the solid curve plots the uncorrected errors* $(\tilde{R}_0, 1-\tilde{R}_1)$*, and the dotted curve plots the corrected errors* $(R_0, 1-R_1)$ *which account for the presence of label noise in the test data.*

measures could be analyzed similarly. Unlike previous theoretical work on this problem, we handle the cases where the supports of P_0 and P_1 may overlap or even be equal, and the noise proportions are asymmetric and unknown.

We also argue that mutual irreducibility is necessary if we require the decontamination operation at population level to satisfy some natural conditions (universality, symmetry, continuity and stability.) Additionally, requiring mutual irreducibility can be equivalently seen as aiming at maximum denoising of the contaminated distributions, or maximum separation of the unknown sources P_0, P_1 for given contaminated distributions. Thus, our discrimination rule is universally consistent in the sense that its performance tends to the optimal performance corresponding to the maximally denoised P_0, P_1, regardless of $\tilde{P}_0, \tilde{P}_1$.

Finally, we investigate two practical implementations of MPE, one based on the ROC for the contaminated data, and the other based on class probability estimation for the contaminated data. The ROC method exhibits good accuracy in the label noise setting on three different data sets, including the nuclear particle classification problem that originally motivated this work. Our CPE implementation, on the other hand, still requires further development.

Appendix A: Mixture proportion consistency result

Blanchard et al. [7] establish strong consistency of $\widehat{\kappa}$, that is, convergence almost surely, although the statement of that consistency result requires a slight correc-

tion. In particular, it is necessary to additionally assume that $\log \max(n_0, n_1) = o(\min(n_0, n_1))$ for the argument to hold. Although the focus of that work is almost sure convergence, the proof can be easily modified to establish convergence in probability, and for that type of convergence, the aforementioned qualification on the growth of the sample sizes is not necessary. Since the present work focuses on convergence in probability, our results also require no additional qualification. See [39] for additional details.

Appendix B: Remaining proofs

B.1. Proof of Proposition 1

Proof. First note that under **(A)**, λ is well-defined and nonnegative. Solving for γ we obtain

$$\gamma = \frac{\lambda(1-\pi_0)-\pi_1}{1-\pi_1-\lambda\pi_0}.$$

The denominator in this expression is positive, which can be seen as follows.

$$\begin{aligned}\lambda &= \frac{\pi_1+\gamma(1-\pi_1)}{1-\pi_0+\gamma\pi_0}\\ &< \frac{1-\pi_0+\gamma(1-\pi_1)}{1-\pi_0+\gamma\pi_0}\\ &< \frac{\gamma(1-\pi_1)}{\gamma\pi_0}\\ &= \frac{1-\pi_1}{\pi_0}.\end{aligned}$$

The first inequality follows from **(A)**, while the second follows from the fact that the mapping $t \mapsto (a+t)/(b+t)$ is strictly decreasing in $t \geq 0$ when $a > b$. Here $a = \gamma(1-\pi_1)$ and $b = \gamma\pi_0$.

Therefore,

$$\begin{aligned}\frac{p_1(x)}{p_0(x)} > \gamma &\iff \frac{p_1(x)}{p_0(x)} > \frac{\lambda(1-\pi_0)-\pi_1}{1-\pi_1-\lambda\pi_0}\\ &\iff [1-\pi_1-\lambda\pi_0]p_1(x) > [\lambda(1-\pi_0)-\pi_1]p_0(x)\\ &\iff (1-\pi_1)p_1(x)+\pi_1 p_0(x) > \lambda[(1-\pi_0)p_0(x)+\pi_0 p_1(x)]\\ &\iff \frac{\tilde{p}_1(x)}{\tilde{p}_0(x)} > \lambda.\end{aligned}$$

□

B.2. Proof of Theorem 11

Proof. Let (P_0, P_1) be mutually irreducible and fixed for the rest of the proof. Observe that if conditions (i) and (ii) are satisfied, (P_0, P_1) must be a ϕ-source.

Namely, by (i) (P_0, P_1) belongs to the domain of ϕ; and since (P_0, P_1) are mutually irreducible, the only possible values for $\phi(P_0, P_1)$ are $(0, 0, P_0, P_1)$ and $(1, 1, P_1, P_0)$. In any case, by the symmetry condition (ii), (P_0, P_1) and (P_1, P_0) are both ϕ-sources. Finally, by stability condition (iv), it must be the case that $\phi(P_0, P_1) = (0, 0, P_0, P_1)$.

Let us now denote ϵ^* the supremum of values ϵ such that, for all $(\pi_0, \pi_1) \in [0, 1]^2$ with $\pi_0 + \pi_1 < \epsilon$, (12) is satisfied. Condition (iv) implies $\epsilon^* > 0$. If $\epsilon^* = 1$, this means that ϕ returns the mutually irreducible solution for mutual contamination with arbitrary contamination weights of (P_0, P_1).

We now consider the case where $\epsilon^* < 1$ and will come to a contradiction. Fix arbitrarily $\epsilon \in (\epsilon^*, 1)$. By definition of ϵ^*, there exists (π_0, π_1) such that $\epsilon^* < \pi_0 + \pi_1 < \epsilon$ and (12) is not satisfied.

Let $(\nu_0, \nu_1, P_0^\epsilon, P_1^\epsilon) = \phi(\psi(\pi_0, \pi_1, P_0, P_1))$ be the contamination proportions and sources identified by ϕ for the contamination $\psi(\pi_0, \pi_1, P_0, P_1)$. Since (12) is not satisfied, and the identified sources uniquely determine the associated contamination proportions, it holds that $(P_0^\epsilon, P_1^\epsilon)$ is a ϕ-source distinct from (P_0, P_1). Finally let us denote (η_0, η_1) the contamination weights of $(P_0^\epsilon, P_1^\epsilon)$ in its mutually irreducible decontamination in terms of (P_0, P_1). (Observe that P_0^ϵ and P_1^ϵ both belong to the convex hull of P_0 and P_1; this implies that $(P_0^\epsilon, P_1^\epsilon)$ decontaminate irreducibly either to (P_0, P_1) or to (P_1, P_0). We assume for now the former case and will come back to the latter case later.)

It must hold that $\eta_0 + \eta_1 \geq \epsilon^*$, otherwise we would have (by definition of ϵ^*) $\phi(P_0^\epsilon, P_1^\epsilon) = (\eta_0, \eta_1, P_0, P_1) = \phi(\psi(0, 0, P_0^\epsilon, P_1^\epsilon))$, contradicting (iv) for the source $(P_0^\epsilon, P_1^\epsilon)$. Moreover, straightforward computations and coefficient identification in the unique representation in terms of (P_0, P_1) lead to the relations

$$(\nu_0, \nu_1) = \left(\frac{\pi_0 - \eta_0}{1 - (\eta_0 + \eta_1)}, \frac{\pi_1 - \eta_1}{1 - (\eta_0 + \eta_1)} \right).$$

It follows that

$$\nu_0 + \nu_1 = 1 - \frac{1 - (\pi_0 + \pi_1)}{1 - (\eta_0 + \eta_1)} \leq 1 - \frac{1 - \epsilon}{1 - \epsilon^*}.$$

In the case where $(P_0^\epsilon, P_1^\epsilon)$ decomposes irreducibly to (P_1, P_0), the first equality above still holds when replacing η_i by $(1 - \eta_i)$. We deduce that in that case $\nu_0 + \nu_1 > 1$.

Now consider a sequence $\epsilon_n \searrow \epsilon^*$, and the associated sequences $(\pi_0^{(n)}, \pi_1^{(n)})$ and $(\eta_0^{(n)}, \eta_1^{(n)})$ constructed as above. By compactness, we can extract a subsequence so that $(\pi_0^{(n)}, \pi_1^{(n)})$ converges to some (π_0^*, π_1^*). Then by construction $\pi_0^* + \pi_1^* = \epsilon^* \in (0, 1)$. On the other hand, for all n either $\nu_0^{(n)} + \nu_1^{(n)} \leq 1 - \frac{1-\epsilon_n}{1-\epsilon^*}$ (which gets arbitrarily close to 0) or $\nu_0^{(n)} + \nu_1^{(n)} \geq 1$. This contradicts the continuity assumption (iii) at point (π_0^*, π_1^*), since by definition of ϵ^* and (iii), it should hold $\phi_\pi(\psi(\pi_0^*, \pi_1^*, P_0, P_1)) = (\pi_0^*, \pi_1^*)$ and thus we should have $\nu_0^{(n)} + \nu_1^{(n)} \to \epsilon^*$.

Conversely, if ϕ is the mutually irreducible decontamination operator, it satisfies (i)–(iv), and is therefore the only decontamination operator having these properties. □

B.3. Proof of Theorem 12

Proof. By Lemmas 2 and 3, feasible quadruples (π_0, π_1, P_0, P_1) for decompositions (1)–(2) under condition **(A)** are in one-to-one correspondence with feasible quadruples $(\tilde{\pi}_0, \tilde{\pi}_1, P_0, P_1)$ for decompositions (6)–(7).

Define $\tilde{\pi}_0^* := \kappa^*(\tilde{P}_1|\tilde{P}_0)$. Proposition 5 applied to (6) easily implies that for any value $\tilde{\pi}_0 \in [0, \tilde{\pi}_0^*]$, there exists a unique P_0 such that $(\tilde{\pi}_0, P_0)$ satisfies (6); also, the solution $(\tilde{\pi}_0^*, P_0^*)$ corresponding to the maximal feasible value of $\tilde{\pi}_0$ is the unique one satisfying **(B)**. A similar conclusion is valid concerning solutions of (7).

Therefore, the feasible region R for proportions (π_0, π_1) in the original model (1)–(2) is obtained as the image of the rectangle $[0, \tilde{\pi}_0^*] \times [0, \tilde{\pi}_1^*]$ via the above one-to-one correspondence. Using the explicit expression for $(\tilde{\pi}_1, \tilde{\pi}_0)$ of Lemma 2, the constraints (14) simply translate the equivalent constraints $\tilde{\pi}_0 \leq \tilde{\pi}_0^*$, $\tilde{\pi}_1 \leq \tilde{\pi}_1^*$.

Since by Lemma 9, under the assumption **(A)** conditions **(B)** and **(C)** are equivalent, then again via the above correspondence, we get existence and unicity of $(\pi_0^*, \pi_1^*, P_0^*, P_1^*)$ for the original formulation (1)–(2), under condition **(C)**. The explicit expression (13) for (π_0^*, π_1^*) is obtained via Lemma 3.

The equality $\pi_0 + \pi_1 = 1 - \frac{(1-\tilde{\pi}_1)(1-\tilde{\pi}_0)}{1-\tilde{\pi}_1\tilde{\pi}_0}$ implies that $\pi_0 + \pi_1$ is a monotone (strictly) increasing function of $\tilde{\pi}_1$ and $\tilde{\pi}_0$. Therefore, the maximum of $\pi_0 + \pi_1$ can only be reached when both $(\tilde{\pi}_1, \tilde{\pi}_0)$ take their maximum value. Since the latter values are attained for the unique feasible quadruple $(\tilde{\pi}_0^*, \tilde{\pi}_1^*, P_0^*, P_1^*)$ in the decoupled problem, the corresponding maximum of $\pi_0 + \pi_1$ for the original formulation is also uniquely attained for the quadruple $(\pi_0^*, \pi_1^*, P_0^*, P_1^*)$.

Finally, by subtracting (1) from (2), we obtain the relation

$$(P_1 - P_0) = (1 - \pi_0 - \pi_1)^{-1}(\tilde{P}_1 - \tilde{P}_0)$$

implying

$$\|P_1 - P_0\|_{TV} = (1 - \pi_0 - \pi_1)^{-1} \left\|\tilde{P}_1 - \tilde{P}_0\right\|_{TV}.$$

Therefore, the maximum (over Λ) of the total variation distance $\|P_1 - P_0\|_{TV}$ is precisely attained for the maximum value of $(\pi_0 + \pi_1)$, and hence corresponds to the unique mutually irreducible solution. □

B.4. Alternate proof of density ratio formula for κ^*

Proposition 23. *Assume that the ROC of the likelihood ratio tests*

$$x \mapsto \mathbf{1}_{\{f(x)/h(x) > \gamma\}}$$

is left-differentiable at $(1, 1)$. *Then* $\kappa^*(F|H)$ *is the slope (left-derivative) of the ROC at* $(1, 1)$.

Proof. The slope of the ROC of an LRT with threshold γ is equal to γ wherever the slope is well defined [33, 37]. The right end-point of the ROC corresponds to $\gamma^* = \operatorname{ess\,inf}_{x \in \operatorname{supp}(H)} \frac{f(x)}{h(x)}$. That is, for all $\gamma > \gamma^*$, the Type I error of the LRT is strictly less than 1, whereas it equals 1 at γ^*. □

B.5. Proof of Theorem 14

We begin by establishing (19) without the absolute value, which is the more challenging direction. The reverse direction will follow easily by the first part of Theorem 13.

By **(D)**, there exists a distribution G and $\gamma \in [0,1]$ such that $F = (1-\gamma)G + \gamma H$ and $\mathrm{supp}(H) \not\subset \mathrm{supp}(G)$. Then G is irreducible with respect to H, and Corollary 7 implies that $\gamma = \kappa^*$. By **(AP2)**, there exists $j \geq 1$ and $S \in \mathcal{S}_j$ such that $G(S) = 0$ and $H(S) > 0$. But then

$$\frac{F(S)}{H(S)} = (1-\gamma)\frac{G(S)}{H(S)} + \gamma = \kappa^*.$$

By the VC inequality and union bound, we have that with probability at least $1 - 2(\frac{1}{n_0} + \frac{1}{n_1})$,

$$\widehat{\kappa} \leq \frac{F(S) + 2\epsilon_1(j, j^{-2}/n_1)}{(H(S) - 2\epsilon_0(j, j^{-2}/n_0))_+} \leq \frac{F(S) + \epsilon}{(H(S) - \epsilon)_+}$$

where $\epsilon := 2(\epsilon_1(j, j^{-2}/n_1) + \epsilon_0(j, j^{-2}/n_0))$. Now let ν be such that $\epsilon = \frac{\nu}{1+\nu} H(S)$, which is achieved by $\nu = \frac{\epsilon}{H(S)-\epsilon}$. Let N be such that $n_0, n_1 \geq N$ implies $\epsilon \leq \frac{1}{2} H(S)$. Then, for $n_0, n_1 \geq N$ and with probability at least $1 - 2(\frac{1}{n_0} + \frac{1}{n_1})$,

$$\begin{aligned}
\widehat{\kappa} &\leq (1+\nu)\frac{F(S) + \epsilon}{H(S)} \\
&= (1+\nu)\kappa^* + \nu \\
&\leq \kappa^* + 2\nu \\
&\leq \kappa^* + \frac{4}{H(S)}\epsilon.
\end{aligned}$$

This establishes the existence of a constant C such that for $n_0, n_1 \geq N$,

$$\Pr\left(\widehat{\kappa} - \kappa^* \geq C\left[\sqrt{\frac{\log n_0}{n_0}} + \sqrt{\frac{\log n_1}{n_1}}\right]\right) \leq \frac{2}{n_0} + \frac{2}{n_1}.$$

The same inequality holds with the absolute value by the first part of Theorem 13, which holds on the same event (samples where the VC bounds hold for all $k \geq 1$) as was used to establish the above inequality.

B.6. Proof of Theorem 18

By Corollary 16, it suffices to show $R_{\tilde{P}, L_\alpha}(\widehat{f}_n) - R^*_{\tilde{P}, L_\alpha} \to 0$ in probability. Toward this end we employ Rademacher complexity analysis. In particular, we will leverage the following result.

Theorem 24. *Let $Z, Z_1, \ldots, Z_n$ be iid random variables taking values in a set $\mathcal{Z}$. Let $\sigma_1, \ldots, \sigma_n$ be iid Rademacher random variables, independent of $Z, Z_1, \ldots, Z_n$. Consider a set of functions $\mathcal{G} \subseteq [a,b]^{\mathcal{Z}}$. $\forall \delta > 0$, with probability $\geq 1-\delta$ with respect to the draw of $Z_1, \ldots, Z_n$, we have*

$$\forall g \in \mathcal{G}, \quad \left| \mathbb{E}[g(Z)] - \frac{1}{n}\sum_{i=1}^{n} g(Z_i) \right| \leq 2\mathfrak{R}_n(\mathcal{G}) + (b-a)\sqrt{\frac{\log 2/\delta}{2n}}, \tag{30}$$

where

$$\mathfrak{R}_n(\mathcal{G}) = \mathop{\mathbb{E}}_{\substack{Z_1,\ldots,Z_n \\ \sigma_1,\ldots,\sigma_n}} \left[\sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i g(Z_i) \right]$$

is the Rademacher complexity of $\mathcal{G}$.

A proof may be found in Mohri et al. [31, Thm. 3.1].

For any $f \in \mathcal{H}$ and loss L', denote the empirical L'-risk

$$\widehat{R}_{L'}(f) := \frac{1}{n}\sum_{i=1}^{n} L'(\tilde{Y}_i, f(X_i)),$$

and denote the objective function $J(f) := \widehat{R}_{L_{\hat{\alpha}}}(f) + \lambda_n \|f\|^2$. Also define $C_0 := \max\{L(0,0), L(1,0)\}$. Observe that $J(\widehat{f}_n) \leq J(0) \leq C_0$. Therefore $\lambda_n \|\widehat{f}_n\|^2 \leq C_0 - \widehat{R}_{L_{\hat{\alpha}}}(\widehat{f}_n) \leq C_0$, and we deduce that $\widehat{f}_n \in B_{\mathcal{H}}(M_n)$, the ball of radius M_n in $\mathcal{H}$, where $M_n := \sqrt{C_0/\lambda_n}$.

Let $\epsilon > 0$, and let $f_\epsilon \in \mathcal{H}$ be such that $R_{\tilde{P},L_\alpha}(f_\epsilon) < R^*_{\tilde{P},L_\alpha} + \frac{\epsilon}{2}$, which is possible since the the reproducing kernel associated with $\mathcal{H}$ is universal [43]. Then

$$\begin{aligned} R_{\tilde{P},L_\alpha}(\widehat{f}_n) - R_{\tilde{P},L_\alpha}(f_\epsilon) = {} & R_{\tilde{P},L_\alpha}(\widehat{f}_n) - \widehat{R}_{L_\alpha}(\widehat{f}_n) \\ & + \widehat{R}_{L_\alpha}(\widehat{f}_n) - \widehat{R}_{L_{\hat{\alpha}}}(\widehat{f}_n) \\ & + \widehat{R}_{L_{\hat{\alpha}}}(\widehat{f}_n) - \widehat{R}_{L_{\hat{\alpha}}}(f_\epsilon) \\ & + \widehat{R}_{L_{\hat{\alpha}}}(f_\epsilon) - \widehat{R}_{L_\alpha}(f_\epsilon) \\ & + \widehat{R}_{L_\alpha}(f_\epsilon) - R_{\tilde{P},L_\alpha}(f_\epsilon). \end{aligned}$$

The first term can be bounded, with probability at least $1 - 1/n$, by

$$\frac{2DBM_n}{\sqrt{n}} + (C_0 + DBM_n)\sqrt{\frac{\ln 2n}{2n}}$$

using the Rademacher complexity bound applied to the class of functions $\mathcal{G} = \{(x,\tilde{y}) \mapsto L(\tilde{y}, f(x)), f \in B_{\mathcal{H}}(M_n)\}$, where $B_{\mathcal{H}}(M_n)$ is the ball of radius M_n (centered at the origin) in $\mathcal{H}$. By the Lipschitz composition property of Rademacher complexity [31, Lemma 4.2], $\mathfrak{R}_n(\mathcal{G}) \leq D\mathfrak{R}_n(B_{\mathcal{H}}(M_n))$. The Rademacher

complexity of $B_{\mathcal{H}}(M_n)$ is further bounded by $BM_n/\sqrt{n}$ [31, Thm 5.5], which gives the first term on the RHS. The second term comes from the observation that functions in $\mathcal{G}$ have ranges confined to $[0, C_0 + DBM_n]$. To see this, recall that losses are by definition nonnegative, that L is Lipschitz in its second argument, and that for any $f \in B_{\mathcal{H}}(M_n)$, we have $\|f\|_\infty = \sup_{x\in\mathcal{X}} |\langle f, k(\cdot, x)\rangle| \leq BM_n$ by the reproducing property and Cauchy-Schwarz.

The fifth term is bounded similarly, with the only additional observation being that $f_\epsilon \in B_{\mathcal{H}}(M_n)$ for n sufficiently large.

The middle term can be bounded by $\lambda_n \|f_\epsilon\|^2$, which tends to zero as $n \to \infty$. This follows from the definition of $\widehat{f}_n$, since $J(\widehat{f}_n) \leq J(f_\epsilon)$ implies $\widehat{R}_{L_{\widehat{\alpha}}}(\widehat{f}_n) - \widehat{R}_{L_{\widehat{\alpha}}}(f_\epsilon) \leq \lambda_n \|f_\epsilon\|^2 - \lambda_n \|\widehat{f}_n\|^2 \leq \lambda_n \|f_\epsilon\|^2$.

To bound the second term, observe that for any $f \in B_{\mathcal{H}}(M_n)$,

$$\begin{aligned}
\widehat{R}_{L_\alpha}(f) - \widehat{R}_{L_{\widehat{\alpha}}}(f) = \frac{1}{n}\Bigg[& \sum_{i:\tilde{Y}_i=1} (\widehat{\alpha} - \alpha) L(1, f(X_i)) \\
& + \sum_{i:\tilde{Y}_i=0} (\alpha - \widehat{\alpha}) L(0, f(X_i)) \Bigg] \\
& \leq |\widehat{\alpha} - \alpha| \sup_{x,\tilde{y}} L(\tilde{y}, f(x)) \\
& \leq |\widehat{\alpha} - \alpha| \left(C_0 + D\|f\|_\infty\right),
\end{aligned}$$

where D is the Lipschitz constant of L. By Cauchy-Schwarz and the reproducing property,

$$\|f\|_\infty = \sup_x |\langle f, k(\cdot, x)\rangle| \leq \|f\|_{\mathcal{H}} B$$

where B is the bound on the kernel. Now $\|f\|_{\mathcal{H}} \leq \sqrt{\frac{C_0}{\lambda_n}}$, and so for the second term to go to zero, we need $|\widehat{\alpha} - \alpha|/\lambda_n$ to go to zero. Under **(A')** and **(C')**, we know that $|\widehat{\alpha} - \alpha|$ converges at a rate of $\sqrt{\frac{\log n}{n}}$, and by our assumption on the rate of decay of λ_n, $|\widehat{\alpha} - \alpha|/\lambda_n$ tends to zero as $n \to \infty$, except on a vanishingly small event.

The fourth term is handled in a similar manner, where again we observe that $f_\epsilon \in B_{\mathcal{H}}(M_n)$ for n sufficiently large.

In summary, we have shown that $R_{\tilde{P}, L_\alpha}(\widehat{f}_n) - R^*_{\tilde{P}, L_\alpha} \leq \epsilon$ with probability tending to one as n (and with it n_0 and n_1) tends to infinity. This concludes the proof.

B.7. Proof of Theorem 19

We start by establishing that L_α is T-clippable and its clipped version is Lipschitz and bounded with constants independent of $\alpha \in (0, 1)$. The loss L being T-clippable implies by definition that both $L_0(t) = L(0, t)$ and $L_1(t) = L(1, t)$

are clippable. Therefore, $L_\alpha(y,t) = (1-\alpha)\mathbf{1}_{\{y=1\}}L_1(t) + \alpha\mathbf{1}_{\{y=0\}}L_0(t)$ is T-clippable (regardless of $\alpha \in (0,1)$.) Denote

$$\widetilde{L}_\alpha(y,t) := L_\alpha(y, \mathrm{Clip}_T(t)) = (1-\alpha)\mathbf{1}_{\{y=1\}}L_1(\mathrm{Clip}_T(t)) + \alpha\mathbf{1}_{\{y=0\}}L_0(\mathrm{Clip}_T(t))\,,$$

and define $C_0 := \max\{L(0,0), L(1,0)\}$. Since L is assumed Lipschitz with constant D, both L_1 and L_0 are Lipschitz and since Clip_T is 1-Lipschitz, by composition $\widetilde{L}_\alpha$ is also a Lipschitz loss (with the same constant D, regardless of $\alpha \in (0,1)$.) Furthermore, since $\mathrm{Clip}_T(t) \in [-T,T]$, we have for all (y,t) and α:

$$\left|\widetilde{L}_\alpha(y,t)\right| \leq \max_{t\in[-T,T]} \max\left(L_0(t), L_1(t)\right) \leq C_0 + DT\,.$$

We proceed to proving the main claim. By Corollary 16, it suffices to show $R_{\tilde{P},L_\alpha}(\check{f}_n) - R^*_{\tilde{P},L_\alpha} \to 0$ in probability. For any f and loss L', denote by $\widehat{R}_{L'}(f)$ the empirical L'-risk of f. Denote the objective function $J(f) := \widehat{R}_{L_{\hat{\alpha}}}(f) + \lambda_n\|f\|^2$. Observe that $J(\widehat{f}_n) \leq J(0) \leq C_0$. Therefore $\lambda_n\|\widehat{f}_n\|^2 \leq C_0 - \widehat{R}_{L_{\hat{\alpha}}}(\widehat{f}_n) \leq C_0$, and we deduce that $\widehat{f}_n \in B_{\mathcal{H}}(M_n)$, the ball of radius M_n in $\mathcal{H}$, where $M_n := \sqrt{C_0/\lambda_n}$.

Let $\epsilon > 0$, and let $f_\epsilon \in \mathcal{H}$ be such that $R_{\tilde{P},L_\alpha}(f_\epsilon) < R^*_{\tilde{P},L_\alpha} + \frac{\epsilon}{2}$, which is possible since the the reproducing kernel associated with $\mathcal{H}$ is universal [43]. We have

$$\begin{aligned}
R_{\tilde{P},L_\alpha}(\check{f}_n) - R_{\tilde{P},L_\alpha}(f_\epsilon) &= R_{\tilde{P},\widetilde{L}_\alpha}(\widehat{f}_n) - R_{\tilde{P},L_\alpha}(f_\epsilon) \\
&= R_{\tilde{P},\widetilde{L}_\alpha}(\widehat{f}_n) - \widehat{R}_{\widetilde{L}_\alpha}(\widehat{f}_n) \\
&\quad + \widehat{R}_{\widetilde{L}_\alpha}(\widehat{f}_n) - \widehat{R}_{\widetilde{L}_{\hat{\alpha}}}(\widehat{f}_n) \\
&\quad + \widehat{R}_{\widetilde{L}_{\hat{\alpha}}}(\widehat{f}_n) - \widehat{R}_{L_{\hat{\alpha}}}(\widehat{f}_n) \\
&\quad + \widehat{R}_{L_{\hat{\alpha}}}(\widehat{f}_n) - \widehat{R}_{L_{\hat{\alpha}}}(f_\epsilon) \\
&\quad + \widehat{R}_{L_{\hat{\alpha}}}(f_\epsilon) - \widehat{R}_{L_\alpha}(f_\epsilon) \\
&\quad + \widehat{R}_{L_\alpha}(f_\epsilon) - R_{\tilde{P},L_\alpha}(f_\epsilon).
\end{aligned}$$

The first and last terms can be bounded, with probability at least $1 - 1/n$, by

$$\frac{2DBM_n}{\sqrt{n}} + (C_0 + D\max(T, B\,\|f_\epsilon\|_\infty))\sqrt{\frac{\ln 4n}{2n}}$$

using Rademacher complexity analysis as was done in the preceding proof. Here D is the Lipschitz constant for L (and thus also for $\widetilde{L}$) and B is the bound on the kernel. Note that since a different loss is used for the first and last terms, we use a union bound and thus introduce an additional factor in the log term.

The third term equals $\widehat{R}_{L_{\hat{\alpha}}}(\check{f}_n) - \widehat{R}_{L_{\hat{\alpha}}}(\widehat{f}_n)$ and is nonpositive by definition of an T-clippable loss.

The middle (4th) term can be bounded (as in the preceding proof) by $\lambda_n \|f_\epsilon\|^2$, which tends to zero as $n \to \infty$.

To bound the second term, observe that for any f,

$$\begin{aligned}\widehat{R}_{\widetilde{L}_\alpha}(f) - \widehat{R}_{\widetilde{L}_{\widehat{\alpha}}}(f) &= \frac{1}{n}\Bigg[\sum_{i:\widetilde{Y}_i=1} (\widehat{\alpha}-\alpha)\widetilde{L}(1, f(X_i)) \\ &\qquad + \sum_{i:\widetilde{Y}_i=0} (\alpha-\widehat{\alpha})\widetilde{L}(0, f(X_i))\Bigg] \\ &\leq |\widehat{\alpha}-\alpha| \sup_{x,\tilde{y}} \widetilde{L}(\tilde{y}, f(x)) \\ &\leq |\widehat{\alpha}-\alpha|(C_0 + DT)\,.\end{aligned}$$

The fifth term is handled in a similar manner, but with the non-clipped loss L instead of $\widetilde{L}$; in this case we have

$$\widehat{R}_{\widetilde{L}_\alpha}(f) - \widehat{R}_{\widetilde{L}_{\widehat{\alpha}}}(f) \leq |\widehat{\alpha}-\alpha|\,(C_0 + D\|f_\epsilon\|_\infty)\,.$$

In summary, we have shown that $R_{\tilde{P},L_\alpha}(\widehat{f}_n) - R^*_{\tilde{P},L_\alpha} \leq \epsilon$ with probability tending to one as n (and with it n_0 and n_1) tends to infinity. This concludes the proof.

Acknowledgements

G. Blanchard was supported in part by the European Community's 7th Framework Programme under the E.U. grant agreement 247022 (MASH Project), and by the DFG grant FOR-1735 (Structural inference in statistics). C. Scott was supported in part by NSF Grants 1422157, 0953135, 1047871, and 1217880.

References

[1] J. M. Adams and G. White. A versatile pulse shape discriminator for charged particle separation and its application to fast neutron time-of-flight spectroscopy. *Nuclear Instruments and Methods in Physics Research*, 1978.

[2] D. Aldous and P. Diaconis. Strong uniform times and finite random walks. *Adv. Appl. Math.*, 8(1):69–97, 1987. MR0876954

[3] S. Ambers, M. Flaska, and S. Pozzi. A hybrid pulse shape discrimination technique with enhanced performance at neutron energies below 500 kev. *Nuclear Instruments and Methods in Physics Research A*, 638:116–121, 2011.

[4] D. Angluin and P. Laird. Learning from noisy examples. *Machine Learning*, 2:343–370, 1988.

[5] J. Aslam and S. Decatur. On the sample complexity of noise-tolerant learning. *Inf. Process. Lett.*, 57:189–195, 1996. MR1379694

[6] P. Bartlett, M. Jordan, and J. McAuliffe. Convexity, classification, and risk bounds. *J. American Statistical Association*, 101(473):138–156, 2006. MR2268032

[7] G. Blanchard, G. Lee, and C. Scott. Semi-supervised novelty detection. *Journal of Machine Learning Research*, 11:2973–3009, 2010. MR2746544

[8] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the 11th Annual Conference on Computational Learning Theory*, pages 92–100, 1998. MR1811574

[9] C. Bouveyron and S. Girard. Robust supervised classification with mixture models: Learning from data with uncertain labels. *Journal of Pattern Recognition*, 42:2649–2658, 2009. MR2509603

[10] C. Brodley and M. Friedl. Identifying mislabeled training data. *Journal of Artifcial Intelligence Research*, 131–167, 1999.

[11] N. H. Bshouty, S. A. Goldman, H. D. Mathias, S. Suri, and H. Tamaki. Noise-tolerant distribution-free learning of general geometric concepts. *J. ACM*, 45(5):863–890, 1998. MR1668163

[12] A. Buja, W. Stuetzle, and Y. Shen. Loss functions for binary class probability estimation and classification: Structure and applications, manuscript, available at `www-stat.wharton.upenn.edu/~buja`, 2005.

[13] N. Cesa-Bianchi, P. Fischer, E. Shamir, and H.-U. Simon. Randomized hypotheses and minimum disagreement hypotheses for learning with noise. In *Proc. Third European Conf. on Computational Learning Theory*, pages 119–133, 1997. MR1476926

[14] V. Denchev, N. Ding, S. V. N. Vishwanathan, and H. Neven. Robust classification with adiabatic quantum optimization. In J. Langford and J. Pineau, editors, *Proc. 29th Int. Conf. on Machine Learning*, pages 863–870, 2012.

[15] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996. MR1383093

[16] N. Ding and S. V. N. Vishwanathan. t-logistic regression. In J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R.S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 514–522. 2010.

[17] B. Frénay and M. Verleysen. Classification in the presence of label noise: A survey. *IEEE Trans. Neural Networks and Learning Systems*, 25:845–869, 2014.

[18] S. Jabbari. PAC-learning with label noise. Master's thesis, University of Alberta, December 2010.

[19] A. Kalai and R. Servedio. Boosting in the presence of noise. *Symposium on Theory of Computing*, pages 196–205, 2003. MR2121046

[20] M. Kearns. Efficient noise-tolerant learning from statistical queries. *Proceedings of the Twenty-Fifth Annual ACM Symposium on Theory of Computing*, pages 392–401, 1993.

[21] O. Koyejo, N. Natarajan, P. Ravikumar, and I. Dhillon. Consistent binary classification with generalized performance metrics. In Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, and K.Q. Weinberger, editors,

Advances in Neural Information Processing Systems 27, pages 2744–2752, 2014.

[22] J. Langford. Tutorial on practical prediction theory for classification. *J. Machine Learning Research*, 6:273–306, 2005. MR2249822

[23] N. Lawrence and B. Schölkopf. Estimating a kernel Fisher discriminant in the presence of label noise. *Proceedings of the International Conference in Machine Learning*, 2001.

[24] E. Lehmann. *Testing Statistical Hypotheses*. Wiley, New York, 1986. MR0852406

[25] T. Liu and D. Tao. Classification with noisy labels by importance reweighting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(3):447–461, 2016.

[26] P. Long and R. Servido. Random classification noise defeats all convex potential boosters. *Machine Learning*, 78:287–304, 2010. MR3108139

[27] N. Manwani and P. S. Sastry. Noise tolerance under risk minimization. *IEEE Trans. on Cybernetics*, 43(3):1146–1151, 2011.

[28] H. Masnadi-Shirazi and N. Vasconcelos. On the design of loss functions for classification: theory, robustness to outliers, and savageboost. In Y. Bengio D. Koller, D. Schuurmans and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 1049–1056. 2009.

[29] L. Mason, J. Baxter, P. Bartlett, and M. Frean. Boosting algorithms as gradient descent. In *Advances in Neural Information Processing Systems 12*, pages 512–518. MIT Press, 2000.

[30] A. Menon, B. Van Rooyen, C. S. Ong, and R. Williamson. Learning from corrupted binary labels via class-probability estimation. In F. Bach and D. Blei, editors, *Proc. 32th Int. Conf. Machine Learning (ICML)*, Lille, France, 2015.

[31] M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of Machine Learning*. MIT Press, 2012. MR3057769

[32] N. Natarajan, I. S. Dhillon, P. Ravikumar, and A. Tewari. Learning with noisy labels. In *Advances in Neural Information Processing Systems 26*, 2013.

[33] W. Peterson, T. Birdsall, and W. Fox. The theory of signal detectability. *Trans. Inst. Radio Engrs., Professional Group on Information Theory*, 4(4):171 –212, 1954. MR0089130

[34] U. Rebbapragada and C. Brodley. Class noise mitigation through instance weighting. *European Conference on Machine Learning*, pages 708–715, 2007.

[35] M. D. Reid and R. C. Williamson. Composite binary losses. *J. Machine Learning Research*, 11:2387–2422, 2010. MR2727769

[36] S. Sabato and N. Tishby. Multi-instance learning with any hypothesis class. *J. Machine Learning Research*, 13:2999–3039, 2012. MR2997719

[37] L. L. Scharf. *Statistical Signal Processing. Detection, Estimation, an Time Series Analysis*. Addison-Wesley, Reading, MA, 1991.

[38] C. Scott. Calibrated asymmetric surrogate losses. *Electronic Journal of Statistics*, 6:958–992, 2012. MR2988435

[39] C. Scott. Notes on weakly supervised learning, 2014. URL web.eecs.umich.edu/~cscott/wsl.pdf.

[40] C. Scott. A rate of convergence for mixture proportion estimation, with application to learning from noisy labels. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2015.

[41] C. Scott, G. Blanchard, and G. Handy. Classification with asymmetric label noise: Consistency and maximal denoising. In *Proc. Conf. on Learning Theory,* JMLR W&CP, volume 30, pages 489–511. 2013.

[42] C. Scott, G. Blanchard, G. Handy, S. Pozzi, and M. Flaska. Classification with asymmetric label noise: Consistency and maximal denoising. Technical Report arXiv:1303.1208, 2013.

[43] I. Steinwart and A. Christmann. *Support Vector Machines*. Springer, 2008. MR2450103

[44] G. Stempfel and L. Ralaivola. Learning SVMs from sloppily labeled data. In *Proc. 19th Int. Conf. on Artificial Neural Networks: Part I*, pages 884–893, 2009.

[45] L. Xu, K. Crammer, and D. Schuurmans. Robust support vector machine training via convex outlier ablation. *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI)*, 2006.

[46] T. Yang, M. Mahdavi, R. Jin, L. Zhang, and Y. Zhou. Multiple kernel learning from noisy labels by stochastic programming. In J. Langford and J. Pineau, editors, *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pages 233–240, New York, NY, USA, 2012. ACM.