JMLR: Workshop and Conference Proceedings vol 30 (2013) 1–23

Classification with Asymmetric Label Noise: Consistency and Maximal Denoising

Clayton Scott CLAYSCOT@UMICH.EDU
Department of Electrical Engineering and Computer Science
University of Michigan
1301 Beal Avenue
Ann Arbor, MI 48109-2122, USA

Gilles Blanchard GILLES.BLANCHARD@MATH.UNI-POTSDAM.DE
Universität Potsdam, Intitut für Mathematik
Am Neuen Palais 10
14469 Potsdam, Germany

Gregory Handy HANDY1@UMBC.EDU
Department of Electrical Engineering and Computer Science
University of Michigan
1301 Beal Avenue
Ann Arbor, MI 48109-2122, USA

Abstract

In many real-world classification problems, the labels of training examples are randomly corrupted. Thus, the set of training examples for each class is contaminated by examples of the other class. Previous theoretical work on this problem 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. We introduce a general framework for classification with label noise that eliminates these assumptions. Instead, we give assumptions ensuring identifiability and the existence of a consistent estimator of the optimal risk, with associated estimation strategies. For any arbitrary pair of contaminated distributions, there is a unique pair of non-contaminated distributions satisfying the proposed assumptions, and we argue that this solution corresponds in a certain sense to maximal denoising. In particular, we find that learning in the presence of label noise is possible even when the class-conditional distributions overlap and the label noise is not symmetric. A key to our approach is a universally consistent estimator of the maximal proportion of one distribution that is present in another, a problem we refer to as "mixture proportion estimation." This work is motivated by a problem in nuclear particle classification.

Keywords: Label noise, consistency, error estimation, mixture proportion estimation

1. Introduction

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

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

where P_0 and P_1, the class-conditional distributions, are probability distributions on a measurable space $(\mathcal{X}, \mathfrak{S})$. The feature vector $X_i^y \in \mathcal{X}$ 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 is when P_0 and P_1 have overlapping support, 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.

This work is motivated by a nuclear particle classification problem that is critical for nuclear nonproliferation. An organic scintillation detector converts the energy of traversing particles to a pulse-shaped waveform that is in turn sampled into a digital signal $X \in \mathbb{R}^d$. The device is sensitive to high-energy neutrons as well as gamma rays, which need to be separated based on their measured pulses, a problem referred to as pulse shape discrimination (PSD) (Adams and White, 1978; Ambers et al., 2011). Unfortunately, even in controlled laboratory settings, it is very difficult to obtain pure samples of neutron and gamma-ray pulses: the fission events that produce neutrons also yield gamma rays in a proportion that is intrinsic to the source material, and cannot be changed. Furthermore, gamma rays are strongly present in background radiation, as well as some neutrons. Thus, PSD is naturally described by the proposed label noise model.

Previous work on classification with random label noise, reviewed below, has not considered the problem in this generality. Our contribution is to introduce general sufficient conditions on the elements P_0, P_1, π_0, π_1 of the contamination models for the existence of a consistent discrimination rule; 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*.

We present a consistent discrimination rule that leverages consistent estimates of the noise proportions. These proportions are recovered in turn via mixture proportion estimation,

which is the problem of estimating the proportion of one distribution present in another, given random samples from both distributions.

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, is non-unique. In particular, 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. Under these conditions, we therefore establish universally consistent learning of (i) a classifier that compensates for everything that could be construed as label noise, and (ii) the corresponding contamination proportions. In particular, we emphasize that the proposed conditions do not put any restrictions on the possible apparent label distributions $\tilde{P}_0, \tilde{P}_1$, so that our consistency result is distribution-free.

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 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 be otherwise only identifiable up to permutation.

1.1. Problem Statement and Notation

We consider the problem of designing a discrimination rule, in the presence of label noise, that is consistent with respect to a given performance measure. To state the problem precisely, we define the following terms. A *classifier* is a measurable function $f : \mathcal{X} \to \{0, 1\}$. A *performance measure* $R(f)$ assigns every classifier to a nonnegative real number, and depends on the true distributions, P_0 and P_1. The optimal performance measure is denoted $R^* = \inf R(f)$, where the infimum is over all classifiers. A *discrimination rule* is a function $\widehat{f}_{m,n} : \mathcal{X}^m \times \mathcal{X}^n \to (\mathcal{X} \to \{0, 1\})$ mapping training data to classifiers. A discrimination rule is *consistent* iff $R(\widehat{f}_{m,n}) \to R^*$ in probability as $\min\{m, n\} \to \infty$.

We focus on the minmax criterion, for which $R(f) = \max\{R_0(f), R_1(f)\}$, where

$$R_0(f) := P_0(f(X) = 1)\,, \qquad\qquad R_1(f) := P_1(f(X) = 0),$$

are the Type I and Type II errors. The optimal performance R^* is called the *minmax* error. This choice of performance measure is primarily for concreteness; we expect no difficulty in

extending our analysis to other performance measures, both frequentist and Bayesian, that can be defined in terms of R_0 and R_1, such as Neyman-Pearson or expected misclassification cost. This is because our approach is grounded on a technique to estimate $R_0(f)$ and $R_1(f)$.

We also introduce the contaminated Type I and II errors:

$$\tilde{R}_0(f) := \tilde{P}_0(f(X) = 1) = (1 - \pi_0)R_0(f) + \pi_0(1 - R_1(f)), \tag{3}$$

$$\tilde{R}_1(f) := \tilde{P}_1(f(X) = 0) = (1 - \pi_1)R_1(f) + \pi_1(1 - R_0(f)). \tag{4}$$

1.2. Related Work

Classification in the presence of label noise has drawn the attention of numerous researchers. 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 (Brodley and Friedl, 1999; Rebbapragada and Brodley, 2007), as well as the use of robust (usually nonconvex) losses (Mason et al., 2000; Xu et al., 2006; Masnadi-Shirazi and Vasconcelos, 2009; Ding and Vishwanathan, 2010; Denchev et al., 2012). 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 (Bouveyron and Girard, 2009). The method of Lawrence and Schölkopf (2001) 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 (2010) and Manwani and Sastry (2011). These works demonstrate a lack of noise tolerance for boosting and empirical risk minimization based on convex losses, respectively, 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 (2009) recently developed a support vector machine with a modified hinge loss. Proper modification of the loss, however, requires knowledge of the noise proportions. Since these proportions are typically not known *a priori*, our consistent estimators of these proportions could make approaches based on convex risk minimization more broadly applicable.

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 (Angluin and Laird, 1988; Kearns, 1993; Aslam and Decatur, 1996; Cesa-Bianchi et al., 1997; Bshouty et al., 1998; Kalai and Servedio, 2003). 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 (1998) under (i), as the basis of co-training. Some new directions and a thorough review of this body of work were recently presented in Jabbari (2010). As we discuss in the next section, new challenges emerge when (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. It is the basis of co-training (Blum and Mitchell, 1998). When $\pi_1 = 0$, we have "one-sided" label noise, and the problem reduces to learning from positive and unlabeled examples, also known as semi-supervised novelty detection; see Blanchard et al. (2010) for a review of this literature. Finally, a basic form of multiple instance learning can be reduced to classification with one-sided label noise (see Sabato and Tishby, 2012).

2. The Challenge of Label Noise

In this section, we overview the challenges posed by label noise. We focus on the population setting ($m, n = \infty$) 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. 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 Blum and Mitchell (1998).

In this section, we assume that P_0 and P_1 are absolutely continuous with respect to a common dominating measure, such as Lebesgue. Let p_0 and p_1 denote corresponding densities. Thus

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

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* $\tilde{p}_0(x) > 0$*,*

$$\frac{p_1(x)}{p_0(x)} > \gamma \iff \frac{\tilde{p}_1(x)}{\tilde{p}_0(x)} > \lambda, \qquad \textit{where} \qquad \lambda := \frac{\pi_1 + \gamma(1 - \pi_1)}{1 - \pi_0 + \gamma\pi_0}. \tag{5}$$

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

Regardless of the performance measure chosen (probability of error, Neyman-Pearson, etc.), the optimal classifier takes the form of a likelihood ratio test (LRT) based on the true densities. 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 Type I and II errors, we will not obtain a classifier that is optimal with respect to the true Type I and II errors, except in very special circumstances. To make this point concrete, we now consider two 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, with some simple algebra it is easy to show that $\lambda = \gamma$ only if $\gamma = 1$. Thus, unless the two classes are equally probable *a priori*, setting the correct λ for the contaminated LRT is not possible, since π_0 and π_1 are unknown.

Minmax. 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 $P_0 = P_1$.

Similar arguments can be made for other criteria, such as Neyman-Pearson. In summary, a classifier that is optimal with respect to the contaminated Type I and II errors is not optimal with respect to the true Type I and II errors, except in special cases. Based on the above discussion, in the setting of asymmetric, random label noise, it is essential to have accurate estimates of true Type I and Type II errors. These estimates, in turn, facilitate the design of discrimination rules with respect to any criterion. For concreteness, in later sections we examine the minmax criterion in detail. However, our approach readily extends to other performance measures that are based on the false positive and negative rates.

3. Alternate Mixture Representation

We introduce an alternative mixture representation that facilitates our subsequent analysis. The following lemma reformulates the problem.

Lemma 1 *Assume (1)-(2) hold. 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 by contraposition 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

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

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

$$\tilde{R}_0(f) = \tilde{P}_0(f(X) = 1) = (1 - \tilde{\pi}_0)R_0(f) + \tilde{\pi}_0(1 - \tilde{R}_1(f)), \tag{8}$$

$$\tilde{R}_1(f) = \tilde{P}_1(f(X) = 0) = (1 - \tilde{\pi}_1)R_1(f) + \tilde{\pi}_1(1 - \tilde{R}_0(f)), \tag{9}$$

where Equations (8) and (9) follow from Lemma 1. By solving for $R_0(f)$ and $R_1(f)$ in (8) and (9), we find

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

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

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. In the next section we address the estimation of $\tilde{\pi}_0$ and $\tilde{\pi}_1$. Note that it is not necessary to estimate π_0 and π_1, although that would be possible in light of Lemma 1.

We conclude this section with a converse to Lemma 1:

Lemma 2 *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 by contraposition 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, for (1) to hold, by matching distributions in a similar way as in the proof of Lemma 1, 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 $\pi_0 = \tilde{\pi}_0(1 - \pi_1)$. 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 1 and 2 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 $0 \leq \tilde{\pi}_1, \tilde{\pi}_0 < 1$).

In the next section, we turn to estimation of $\tilde{\pi}_0, \tilde{\pi}_1$. We also address the 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)? The equivalent representations (6)-(7) are pivotal to answering both questions, because they are *decoupled* in the sense that the unknown distribution P_0 only enters in (6), and P_1 only in (7). Therefore, we can solve (6) and (7) separately, and each of these reduces to a problem of mixture proportion estimation, as we explain next.

4. Mixture Proportion Estimation and Mutual Irreducibility

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

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

where $0 \leq \nu \leq 1$. Mixture proportion estimation is the following problem: given iid training samples $Z_F^m \in \mathcal{X}^m$ and $Z_H^n \in \mathcal{X}^n$ of sizes m and n from F and H respectively, and no information about G, estimate ν. This problem was previously addressed by Blanchard et al. (2010), and here we relate the necessary definitions and results from that work.

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

Definition 1 *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. (2010).

Proposition 2 *Let F, H be probability distributions. If $F \neq H$, there is a unique $\nu^* \in [0, 1)$ and G such that $F = (1-\nu^*)G + \nu^* H$, and such that G is irreducible with respect to H. If we additionally define $\nu^* = 1$ when $F = H$, then in all cases*

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

By this result, the following is well-defined.

Definition 2 *For any two probability distributions F, H, define*

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

Thus, G is irreducible with respect to H if and only if $\nu^*(G, H) = 0$. Additionally, it is not hard to show that for any two distributions F and H, $\nu^*(F, H) = \inf_{A \in \mathfrak{S}} F(A)/H(A)$. Similarly, when F and H have densities f and h, $\nu^*(F, H) = \operatorname{ess\,inf}_{x \in \mathrm{supp}(H)} f(x)/h(x)$. These identities make it possible to check irreducibility in different scenarios. For example, $\nu^*(G, H) = 0$ whenever the support of G does not contain the support of H. Even if the supports are equal, two distributions can be mutually irreducible, as in the case of two Gaussians with distinct means and equal variances.

To consolidate the above notions, we state the following corollary.

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

Blanchard et al. also studied an estimator $\widehat{\nu} = \widehat{\nu}(Z_F^m, Z_H^n)$ of $\nu^*(F, H)$. They show that $\widehat{\nu}$ is strongly universally consistent, i.e., that for any F and H, $\widehat{\nu} \to \nu^*(F, H)$ almost surely. The particular form of the estimator is not important here; only its consistency is relevant for our purposes. See Appendix A for a correction to the statement of the consistency result of Blanchard et al. (2010); this correction does not affect the present analysis.

Lemma 1 allows us to estimate $\tilde{\pi}_0$ and $\tilde{\pi}_1$ using $\widehat{\nu}$. Recalling the result of Lemma 1, the distributions $\tilde{P}_0$ and $\tilde{P}_1$ can be written

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

By Corollary 1, we can estimate $\tilde{\pi}_0$ 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 $\tilde{P}_0$.

To ensure this condition, we now introduce the following identifiability assumption:

(C) P_0 and P_1 are mutually irreducible.

Note that it follows from assumption **(C)** that $P_0 \neq P_1$. We now establish that **(C)** and **(B)** are essentially equivalent.

Lemma 3 *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. ■

To summarize, if **(A)** and **(C)** hold, then we can consistently estimate $\tilde{\pi}_0$ and $\tilde{\pi}_1$, and therefore can also consistently estimate $R_0(f)$ and $R_1(f)$ via Eqns. (10)-(11). These ideas are developed in the next section.

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 2 for the label noise contamination model.

Proposition 3 *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 $\tilde{\pi}_0^* := \nu^*(\tilde{P}_0, \tilde{P}_1) < 1$ and $\tilde{\pi}_1^* := \nu^*(\tilde{P}_1, \tilde{P}_0) < 1$, it holds*
$$\pi_0^* = \frac{\tilde{\pi}_0^*(1-\tilde{\pi}_1^*)}{1-\tilde{\pi}_1^*\tilde{\pi}_0^*}, \qquad \pi_1^* = \frac{\tilde{\pi}_1^*(1-\tilde{\pi}_0^*)}{1-\tilde{\pi}_1^*\tilde{\pi}_0^*}. \tag{12}$$
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*
$$\pi_0 + \pi_1\tilde{\pi}_0^* \leq \tilde{\pi}_0^*, \qquad \pi_1 + \pi_0\tilde{\pi}_1^* \leq \tilde{\pi}_1^*. \tag{13}$$
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* (13) *are active;*
   - *the unique maximizer over Λ of the total variation distance $\|P_0 - P_1\|_{TV}$.*

The proof of the proposition relies on the explicit one-to-one correspondence established in Lemmas 1 and 3 between the solutions of the original decomposition (1)-(2) and its decoupled reformulation (6)-(7). The result of Proposition 2 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 proposition is visualized on Figure 1 (see appendix). In particular, point 1 of the proposition 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.

5. Estimating Type I and Type II Errors

We denote the training data by $Z_0^m = (X_0^1, ..., X_0^m) \in \mathcal{X}^m$, and $Z_1^n = (X_1^1, ..., X_1^n) \in \mathcal{X}^n$. Given a classifier f, and iid samples Z_0^m and Z_1^n, we define the following estimates of the contaminated Type I and Type II errors:

$$\widehat{\tilde{R}}_0(f, Z_0^m) := \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}_{\{f(X_0^i)\neq 0\}}, \qquad \widehat{\tilde{R}}_1(f, Z_1^n) := \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{f(X_1^i)\neq 1\}}.$$

Following the theory developed in Section 4, define the estimates of $\tilde{\pi}_0$ and $\tilde{\pi}_1$ as

$$\widehat{\tilde{\pi}}_0(Z_0^m, Z_1^n) := \widehat{\nu}(Z_0^m, Z_1^n), \qquad \widehat{\tilde{\pi}}_1(Z_0^m, Z_1^n) := \widehat{\nu}(Z_1^n, Z_0^m),$$

where $\widehat{\nu}$ is the estimator of Blanchard et al. (2010).

Plugging these estimates into Equations (10) and (11), we define the following estimates for the Type I and Type II errors:

$$\begin{aligned} \widehat{R}_0(f, Z_0^m, Z_1^n) &:= 1 - \widehat{\tilde{R}}_1(f, Z_1^n) - \frac{1 - \widehat{\tilde{R}}_0(f, Z_0^m) - \widehat{\tilde{R}}_1(f, Z_1^n)}{1 - \widehat{\tilde{\pi}}_0(Z_0^m, Z_1^n)}, \\ \widehat{R}_1(f, Z_0^m, Z_1^n) &:= 1 - \widehat{\tilde{R}}_0(f, Z_0^m) - \frac{1 - \widehat{\tilde{R}}_1(f, Z_1^n) - \widehat{\tilde{R}}_0(f, Z_0^m)}{1 - \widehat{\tilde{\pi}}_1(Z_0^m, Z_1^n)}. \end{aligned} \tag{14}$$

For brevity, we will sometimes write $\widehat{R}_i(f)$. The following theorem shows that the estimators $\widehat{R}_i(f)$ converge uniformly in probability to $R_i(f)$.

Theorem 1 *Let $\{\mathcal{F}_k\}_{k=1}^{\infty}$ denote a family of sets of classifiers, with $\mathcal{F}_k$ having finite VC-dimension V_k. Let $k(m,n)$ take values in $\mathbb{N}$ such that*

$$\frac{V_{k(m,n)} \log(\min(m,n))}{\min(m,n)} \to 0,$$

as $\min(m,n) \to \infty$. If assumptions **(A)** *and* **(C)** *hold, then, as $\min(m,n) \to \infty$,*

$$\sup_{f \in \mathcal{F}_{k(m,n)}} |\widehat{R}_i(f, Z_0^m, Z_1^n) - R_i(f)| \to 0$$

in probability for $i = 0, 1$.

The proof consists of a showing that $\widehat{\tilde{R}}_0(f, Z_0^m)$ and $\widehat{\tilde{R}}_1(f, Z_1^n)$ converge uniformly to $\tilde{R}_0(f)$ and $\tilde{R}_1(f)$ (by the VC inequality), that $\widehat{\tilde{\pi}}_i \to \tilde{\pi}_i$ in probability, $i = 0, 1$ (by the result of Blanchard et al.), and a continuity argument.

In the next section, we use $\widehat{R}_0$ and $\widehat{R}_1$ to develop a consistent minmax classifier. A similar development should be possible for other criteria depending on Type I and II errors.

6. Minmax Consistency

Define the max error of a classifier f as

$$R(f) \quad := \quad \max\{R_0(f), R_1(f)\}. \tag{15}$$

Let $\mathcal{F}$ denote an arbitrary set of classifiers. We define the minmax error over $\mathcal{F}$ as

$$R(\mathcal{F}) \quad := \quad \inf_{f \in \mathcal{F}} R(f).$$

Let $\mathcal{F}_0$ denote the set of all classifiers, and define $R^* := R(\mathcal{F}_0)$, the minmax error. Define the estimates of $R(f)$ and $R(\mathcal{F})$ as

$$\widehat{R}(f) := \max\{\widehat{R}_0(f), \widehat{R}_1(f)\}, \qquad\qquad \widehat{R}(\mathcal{F}) := \inf_{f \in \mathcal{F}} \widehat{R}(f).$$

Now let $\tau_k > 0$ be a sequence such that $\tau_k \to 0$ as $k \to \infty$. Define $\widehat{f}_k$ to be any classifier

$$\widehat{f}_k \quad \in \quad \{f \in \mathcal{F}_k : \widehat{R}(f) \leq \widehat{R}(\mathcal{F}_k) + \tau_k\}. \tag{16}$$

This construction allows us to avoid assuming the existence of an empirical error minimizer.

Let $\{\mathcal{F}_k\}_{k=1}^{\infty}$ denote a family of sets of classifiers. The following universal approximation property is known to be satisfied for various families of VC classes, such as histograms, decision trees, neural networks, and polynomial classifiers.

(D) For all distributions Q and measurable functions $\tilde{f} : \mathcal{X} \to \{0, 1\}$,

$$\lim_{k \to \infty} \inf_{f \in \mathcal{F}_k} Q(f(X) \neq \tilde{f}(X)) = 0.$$

Theorem 1 gives us control over the estimation error. Condition **(D)** provides control of the approximation error.

Lemma 4 *Let $\{\mathcal{F}_k\}_{k=1}^{\infty}$ denote a sequence of classifier sets. If assumption* **(D)** *holds, then*

$$\lim_{k \to \infty} \inf_{f \in \mathcal{F}_k} R(f) = R^*.$$

We can now state the consistency result. This result is comparable in form to a classical consistency result in the standard classification setup, see Theorem 18.1 of Devroye et al. (1996) where a condition similar to **(D)**, or more precisely to Lemma 4, is discussed.

Theorem 2 *Let $\{\mathcal{F}_k\}_{k=1}^{\infty}$ be a family of sets of classifiers, with $\mathcal{F}_k$ having VC-dimension $V_j < \infty$. Let $k(m, n)$ take values in $\mathbb{N}$ such that $k(m, n) \to \infty$ as $\min(m, n) \to \infty$. If*

$$\frac{V_{k(m,n)} \log(\min(m, n))}{\min(m, n)} \to 0,$$

as $\min(m, n) \to \infty$ and assumptions **(A)**, **(C)**, *and* **(D)** *hold, then $R(\widehat{f}_{k(m,n)}) \to R^*$ in probability as $\min(m, n) \to \infty$.*

If conditions **(A)** or **(C)** fail to hold, our discrimination rule is still consistent with respect to the maximally denoised versions of $\tilde{P}_0$ and $\tilde{P}_1$, which always exist and are unique. In this sense, our analysis is distribution free and the consistency is universal.

The proof of Theorem 2 (see appendix for details) proceeds by a decomposition into estimation and approximation errors (denoting $k = k(m, n)$ for brevity),

$$R(\widehat{f}_k) - R^* = R(\widehat{f}_k) - R(\mathcal{F}_k) + R(\mathcal{F}_k) - R^*.$$

The approximation error goes to zero by Lemma 4. The estimation error is bounded as follows. For the sake of argument, assume $R(\mathcal{F}_k)$ is realized by $f_k^* \in \mathcal{F}_k$. Then

$$R(\widehat{f}_k) - R(\mathcal{F}_k) = R(\widehat{f}_k) - R(f_k^*) \leq \widehat{R}(\widehat{f}_k) - \widehat{R}(f_k^*) + \epsilon \leq 2\epsilon,$$

where the first inequality holds for any $\epsilon > 0$, with probability going to one, by Theorem 1. The second inequality holds by definition of $\widehat{f}_k$, for k sufficiently large.

7. Conclusion

We have argued 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 Blanchard et al. (2010) on mixture proportion estimation to design consistent estimators of the false positive and negative probabilities. These estimators are applied to establish a consistent minmax classifier, and it seems clear that other performance measures could be analyzed similarly. Unlike previous theoretical work on this problem, we allow that the supports of P_0 and P_1 may overlap or even be equal, the noise is asymmetric, and that the performance measure is not the probability of error. We also argued that 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.

Acknowledgments

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

Appendix A. Correction to Statement of Consistency Result of Blanchard et al.

Blanchard et al. (2010) establish an estimator $\widehat{\nu}$ that converges to $\nu^*(F, H)$ almost surely for any F and H, where convergence takes place as $\min(m, n) \to \infty$. The statement of that consistency result requires a slight correction. In particular, it is necessary to additionally assume that $\log \max(m, n) = o(\min(m, n))$ 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.

Appendix B. Geometry of solutions of (1)-(2)

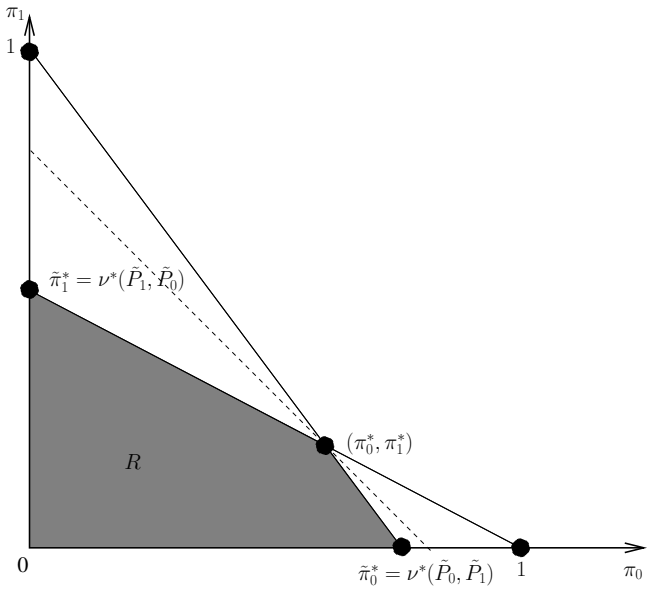

Figure 1: Geometry of the feasible region R for proportions (π_0, π_1) solutions of the contamination model (1)-(2), when contaminated distributions $(\tilde{P}_0, \tilde{P}_1)$ are observed and the true distributions (P_0, P_1) are unknown. Each feasible (π_0, π_1) corresponds to a single associated 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 R.

Appendix C. Remaining Proofs

C.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,

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

■

C.2. Proof of Proposition 3

Proof By Lemmas 1 and 2, 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 $\tilde{\pi}_0^* := \nu^*(\tilde{P}_1, \tilde{P}_0)$. Proposition 2 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 1, the constraints (13) simply translate the equivalent constraints $\tilde{\pi}_0 \leq \tilde{\pi}_0^*$, $\tilde{\pi}_1 \leq \tilde{\pi}_1^*$.

Since by Lemma 3, 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 (12) for (π_0^*, π_1^*) is obtained via Lemma 2.

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

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

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. ■

C.3. Proof of Theorem 1

The following two lemmas allows us to deduce uniform convergence of $\widehat{R}_i$ from uniform convergence of $\widehat{\tilde{R}}_0$ and $\widehat{\tilde{R}}_1$, and consistency of $\widehat{\tilde{\pi}}_0$, and $\widehat{\tilde{\pi}}_1$. They will be used in the proof of Theorem 1.

Lemma 5 *Let $\{\mathcal{F}_j\}_{j=1}^{\infty}$ denote a sequence of classifier sets, with $\mathcal{F}_j$ having finite VC-dimension V_j. Let $k(m,n)$ take values in $\mathbb{N}$ such that*

$$\frac{V_{k(m,n)} \log(\min(m,n))}{\min(m,n)} \to 0, \tag{17}$$

Then

$$\sup_{f \in \mathcal{F}_{k(m,n)}} |\widehat{\tilde{R}}_0(f, Z_0^m) - \tilde{R}_0(f)| \;\; \to \;\; 0,$$

in probability, and

$$\sup_{f \in \mathcal{F}_{k(m,n)}} |\widehat{\tilde{R}}_1(f, Z_1^n) - \tilde{R}_1(f)| \;\; \to \;\; 0$$

in probability.

Proof Let $k = k(m,n)$. We must show that for all $\epsilon > 0$

$$\lim_{\min(m,n)\to\infty} \tilde{P}_0^m(\sup_{f \in \mathcal{F}_k} |\widehat{\tilde{R}}_0(f, Z_0^m) - \tilde{R}_0(f)| > \epsilon) = 0$$

and

$$\lim_{\min(m,n)\to\infty} \tilde{P}_1^n(\sup_{f\in\mathcal{F}_k} |\hat{\tilde{R}}_1(f, Z_1^n) - \tilde{R}_1(f)| > \epsilon) = 0.$$

Let $\ell = \min(m,n)$ and $\epsilon > 0$. By Theorem 12.5 in Devroye et al. (1996), it suffices to show that $8s(\mathcal{F}_k, \ell)e^{-\ell\epsilon^2/32} \to 0$, as $\ell \to \infty$. Theorem 13.3 in Devroye et al. (1996) provides ℓ^{V_k} as an upper bound on the shatter coefficient. Therefore, we have

$$\begin{aligned} 8s(\mathcal{F}_k, \ell)e^{-\ell\epsilon^2/32} &\leq 8\ell^{V_k}e^{-\ell\epsilon^2/32} \\ &= 8e^{-\ell\epsilon^2/32 + V_k\log(\ell)}. \end{aligned}$$

This final term clearly goes to zero by (17). ■

Lemma 6 *(Extension of Continuous Mapping Theorem) Let Q_0, Q_1 be probability distributions. Let $\mathcal{F}_0$ denote the set of all classifiers, and $\{\mathcal{F}_j\}_{j=1}^{\infty}$ denote a family of sets of classifiers. Let $k(m,n)$ take values in $\mathbb{N}$ such that $k(m,n) \to \infty$ as $\min(m,n) \to \infty$. Denote $k = k(m,n)$. Let*

$$\begin{aligned} \widehat{A} &: \mathcal{F}_0 \times \mathcal{X}^m \times \mathcal{X}^n \to \mathbb{R} \\ \widehat{B} &: \mathcal{F}_0 \times \mathcal{X}^m \times \mathcal{X}^n \to \mathbb{R} \\ A &: \mathcal{F}_0 \to \mathbb{R} \\ B &: \mathcal{F}_0 \to \mathbb{R}. \end{aligned}$$

Assume $\sup_{f\in F_k} |\widehat{A}(f, Z_0^m, Z_1^n) - A(f)| \to 0$ and $\sup_{f\in F_k} |\widehat{B}(f, Z_0^m, Z_1^n) - B(f)| \to 0$, in probability, where Z_0^m and Z_1^n are iid random samples governed by the product measures Q_0^m and Q_1^n. If $g : \Omega \subseteq \mathbb{R}\times\mathbb{R} \to \mathbb{R}$ is continuous at $(A(f), B(f))$ for all $f \in \mathcal{F}_0$, then as $\min(m,n) \to \infty$, $\sup_{f\in\mathcal{F}_k} |g(\widehat{A}(f, Z_0^m, Z_1^n), \widehat{B}(f, Z_0^m, Z_1^n)) - g(A(f), B(f))| \to 0$ in probability.

Proof For an arbitrary f and samples of sizes m and n, by the definition of continuity, for all $\epsilon > 0$, there exists a $\delta_\epsilon > 0$ such that

$$\begin{aligned} ||(\widehat{A}(f, Z_0^m, Z_1^n), \widehat{B}(f, Z_0^m, Z_1^n)) - (A(f), B(f))||_2 < 2\delta_\epsilon \\ \implies |g(\widehat{A}(f, Z_0^m, Z_1^n), \widehat{B}(f, Z_0^m, Z_1^n)) - g(A(f), B(f))| < \epsilon. \end{aligned}$$

Since $||\cdot||_1 \geq ||\cdot||_2$, it follows that

$$\begin{aligned} ||(\widehat{A}(f, Z_0^m, Z_1^n), \widehat{B}(f, Z_0^m, Z_1^n)) - (A(f), B(f))||_1 < 2\delta_\epsilon \\ \implies |g(\widehat{A}(f, Z_0^m, Z_1^n), \widehat{B}(f, Z_0^m, Z_1^n)) - g(A(f), B(f))| < \epsilon. \end{aligned}$$

From this, we can conclude that

$$\begin{aligned} \sup_{f\in\mathcal{F}_k} ||(\widehat{A}(f, Z_0^m, Z_1^n), \widehat{B}(f, Z_0^m, Z_1^n)) - (A(f), B(f))||_1 < 2\delta_\epsilon \\ \implies \sup_{f\in\mathcal{F}_k} |g(\widehat{A}(f, Z_0^m, Z_1^n), \widehat{B}(f, Z_0^m, Z_1^n))) - g(A(f), B(f))| \leq \epsilon. \end{aligned}$$

for all m, n. Now,

$$\begin{aligned} 0 &\leq Q_0^m \otimes Q_1^n(\sup_{f\in\mathcal{F}_k} ||(\widehat{A}(f,Z_0^m,Z_1^n),\widehat{B}(f,Z_0^m,Z_1^n)) - (A(f),B(f))||_1 > 2\delta_\epsilon) \\ &= Q_0^m \otimes Q_1^n(\sup_{f\in\mathcal{F}_k} |\widehat{A}(f,Z_0^m,Z_1^n) - A(f)| + |\widehat{B}(f,Z_0^m,Z_1^n) - B(f)| > 2\delta_\epsilon) \\ &\leq Q_0^m \otimes Q_1^n(\sup_{f\in\mathcal{F}_k} |\widehat{A}(f,Z_0^m,Z_1^n) - A(f)| > \delta_\epsilon) \\ &\quad + Q_0^m \otimes Q_1^n(\sup_{f\in\mathcal{F}_k} |\widehat{B}(f,Z_0^m,Z_1^n) - B(f)| > \delta_\epsilon). \end{aligned}$$

Taking the limit as $\min(m,n)\to\infty$ takes the last inequality to 0, based on our assumption of convergence in probability. Therefore, we have

$$\lim_{\min(m,n)\to\infty} Q_0^m \otimes Q_1^n(\sup_{f\in\mathcal{F}_k} ||(\widehat{A}(f,Z_0^m,Z_1^n),\widehat{B}(f,Z_0^m,Z_1^n)) - (A(f),B(f))||_1 < 2\delta_\epsilon) = 1. \tag{18}$$

It follows from a previous implication that

$$\begin{aligned} &Q_0^m \otimes Q_1^n(\sup_{f\in\mathcal{F}_k} ||(\widehat{A}(f,Z_0^m,Z_1^n),\widehat{B}(f,Z_0^m,Z_1^n)) - (A(f),B(f))||_1 < 2\delta_\epsilon) \\ &\qquad \leq Q_0^m \otimes Q_1^n(\sup_{f\in\mathcal{F}_k} |g(\widehat{A}(f,Z_0^m,Z_1^n),\widehat{B}(f,Z_0^m,Z_1^n)) - g(A(f),B(f))| \leq \epsilon). \end{aligned}$$

Combining this inequality with equation (18) yields,

$$\lim_{\min(m,n)\to\infty} Q_0^m \otimes Q_1^n(\sup_{f\in\mathcal{F}_k} |g(\widehat{A}(f,Z_0^m,Z_1^n),\widehat{B}(f,Z_0^m,Z_1^n))) - g(A(f),B(f))| \leq \epsilon) = 1,$$

and the result follows. ■

We will prove the theorem for $i = 0$, the other case being similar.

Proof Let $k = k(m,n)$ for brevity. Substituting equations (7) and (9) into the following subtraction yields

$$\begin{aligned} \widehat{R}_0(f,Z_0^m,Z_1^n) - R_0(f) &= 1 - \widehat{\tilde{R}}_1(f,Z_1^n) - \frac{1-\widehat{\tilde{R}}_0(f,Z_0^m) - \widehat{\tilde{R}}_1(f,Z_1^n)}{1-\widehat{\tilde{\pi}}_0(Z_0^m,Z_1^n)} \\ &\quad -(1-\tilde{R}_1(f) - \frac{1-\tilde{R}_0(f)-\tilde{R}_1(f)}{1-\tilde{\pi}_0}) \\ &= \tilde{R}_1(f) - \widehat{\tilde{R}}_1(f,Z_1^n) - \frac{1-\widehat{\tilde{R}}_0(f,Z_0^m) - \widehat{\tilde{R}}_1(f,Z_1^n)}{1-\widehat{\tilde{\pi}}_0(Z_0^m,Z_1^n)} \\ &\quad + \frac{1-\tilde{R}_0(f)-\tilde{R}_1(f)}{1-\tilde{\pi}_0}. \end{aligned}$$

Take $\epsilon > 0$. By Lemma 5, we have that

$$L_1 := \lim_{\min(m,n)\to\infty} \tilde{P}_1^n(\sup_{f\in\mathcal{F}_k} |\tilde{R}_1(f) - \widehat{\tilde{R}}_1(f,Z_1^n)| > \frac{\epsilon}{3}) = 0.$$

Now consider the following,

$$\sup_{f\in\mathcal{F}_k}\Big|\frac{1-\tilde{R}_0(f)}{1-\tilde{\pi}_0}-\frac{1-\widehat{\tilde{R}}_0(f,Z_0^m)}{1-\widehat{\tilde{\pi}}_0(Z_0^m,Z_1^n)}\Big|=\sup_{f\in F_k}|h(\tilde{R}_0(f),\tilde{\pi}_0)-h(\widehat{\tilde{R}}_0(f,Z_0^m),\widehat{\tilde{\pi}}_0(Z_0^m,Z_1^n))|,$$

where $h(x,y)=(1-x)/(1-y)$. This function is continuous on $\Omega=\mathbb{R}\times(\mathbb{R}\backslash\{0\})$. By **(A)** and Lemma 1, we have that $\tilde{\pi}_0<1$ and therefore this function is continuous at $(\tilde{R}_0(f),\tilde{\pi}_0)$. By **(C)**, $\tilde{\pi}_0=\nu^*(\tilde{P}_0,\tilde{P}_1)$. Furthermore, Theorem 8 of Blanchard et al. (2010) implies that $\widehat{\tilde{\pi}}_0(Z_0^m,Z_1^n)$ converges in probability to $\tilde{\pi}_0$, and by Lemma 5, we have that

$$\sup_{f\in\mathcal{F}_k}|\widehat{\tilde{R}}_0(f,Z_0^m)-\tilde{R}_0(f)|\to 0,$$

in probability. Thus, the conditions of Lemma 6 are met with $\widehat{A}(f,Z_0^m,Z_1^n)=\widehat{\tilde{R}}_0(f,Z_0^m)$, and $\widehat{B}(f,Z_0^m,Z_1^n)=\widehat{\tilde{\pi}}_0(Z_0^m,Z_1^n)$. By applying Lemma 6, we conclude that

$$\lim_{\min(m,n)\to\infty}\tilde{P}_0^m\otimes\tilde{P}_1^n(\sup_{f\in F_k}|h(\tilde{R}_0(f),\tilde{\pi}_0)-h(\widehat{\tilde{R}}_0(f,Z_0^m),\widehat{\tilde{\pi}}_0(Z_0^m,Z_1^n))|>\frac{\epsilon}{3})=0.$$

So we now define

$$L_2\quad:=\quad\lim_{\min(m,n)\to\infty}\tilde{P}_0^m\otimes\tilde{P}_1^n(\sup_{f\in\mathcal{F}_k}\Big|\frac{1-\tilde{R}_0(f)}{1-\tilde{\pi}_0}-\frac{1-\widehat{\tilde{R}}_0(f,Z_0^m)}{1-\widehat{\tilde{\pi}}_0(Z_0^m,Z_1^n)}\Big|>\frac{\epsilon}{3})=0.$$

A similar argument can be made to show that

$$L_3\quad:=\quad\lim_{\min(m,n)\to\infty}\tilde{P}_0^m\otimes\tilde{P}_1^n(\sup_{f\in\mathcal{F}_k}\Big|\frac{-\tilde{R}_1(f)}{1-\tilde{\pi}_0}-\frac{-\widehat{\tilde{R}}_1(f,Z_1^n)}{1-\widehat{\tilde{\pi}}_0(Z_0^m,Z_1^n)}\Big|>\frac{\epsilon}{3})=0.$$

We conclude the proof by applying the triangle inequality,

$$\begin{aligned}\lim_{\min(m,n)\to\infty}\tilde{P}_0^m\otimes\tilde{P}_1^n(\sup_{f\in\mathcal{F}_k}|\widehat{R}_0(f,Z_0^m,Z_1^n)-R_0(f)|>\epsilon)\quad&\leq\quad L_1+L_2+L_3\\&=\quad 0.\end{aligned}$$

■

C.4. Proof of Lemma 4

Proof Let $\epsilon>0$ and let $\tilde{f}\in\mathcal{F}_0$ be a measurable function such that $R(\tilde{f})\leq R^*+\frac{\epsilon}{2}$. Also let $\wedge$ and $\vee$ denote logical "and" and "or". Take $\tilde{P}=\frac{1}{2}P_0+\frac{1}{2}P_1$. By assumption **(D)**, there exists a $k_0\in\mathbb{N}$, such that for every $k\geq k_0$ there exists a $f\in\mathcal{F}_k$ such that

$$\tilde{P}(f(X)\neq\tilde{f}(X))<\frac{\epsilon}{4}.$$

Combining this with the definition of $\tilde{P}$ yields, for such f,

$$\begin{aligned}P_0(f(X)\neq\tilde{f}(X))\quad&\leq\quad 2\tilde{P}(f(X)\neq\tilde{f}(X))\\&<\quad\frac{\epsilon}{2}.\end{aligned}$$

Therefore, for all $k \geq k_0$, there exists a $f \in \mathcal{F}_k$ such that

$$\begin{aligned}
\frac{\epsilon}{2} &> P_0(f(X) \neq \tilde{f}(X)) \\
&= P_0((f(X) = 1 \wedge \tilde{f}(X) = 0) \vee (f(X) = 0 \wedge \tilde{f}(X) = 1)) \\
&= P_0(f(X) = 1 \wedge \tilde{f}(X) = 0) + P_0(f(X) = 0 \wedge \tilde{f}(X) = 1) \\
&\geq P_0(f(X) = 1 \wedge \tilde{f}(X) = 0) - P_0(f(X) = 0 \wedge \tilde{f}(X) = 1) \\
&= P_0(f(X) = 1) - P_0(\tilde{f}(X) = 1) \\
&= R_0(f) - R_0(\tilde{f}).
\end{aligned}$$

In the same manner, it can be shown that $\epsilon/2 > R_1(f) - R_1(f^*)$ for the same $f \in \mathcal{F}_k$. This establishes the existence for all $k \geq k_0$ of a $f \in \mathcal{F}_k$ such that

$$\begin{aligned}
R(f) = \max\{R_0(f), R_1(f)\} &\leq \max\{R_0(\tilde{f}), R_1(\tilde{f})\} + \frac{\epsilon}{2} \\
&= R(\tilde{f}) + \frac{\epsilon}{2} \\
&\leq R^* + \epsilon.
\end{aligned}$$

Since ϵ was arbitrary the result now follows. ■

C.5. Proof of Theorem 2

Proof Let $\epsilon > 0$, $\delta > 0$, and $k = k(m, n)$. We need to show that for m, n sufficiently large,

$$\tilde{P}_0^m \otimes \tilde{P}_1^n(R(\widehat{f}_k) - R^* < \epsilon) > 1 - \delta.$$

Consider the decomposition

$$R(\widehat{f}_k) - R^* = R(\widehat{f}_k) - R(\mathcal{F}_k) + R(\mathcal{F}_k) - R^*.$$

Lemma 4 implies that for m and n significantly large, $R(\mathcal{F}_k) - R^* < \epsilon/2$. We will now bound the $R(\widehat{f}_k) - R(\mathcal{F}_k)$ term. By the definition of $R(\mathcal{F}_k)$, there exists $f_k^* \in \mathcal{F}_k$ such that $R(f_k^*) \leq R(\mathcal{F}_k) + \epsilon/8$. It follows that

$$\begin{aligned}
R(\widehat{f}_k) - R(\mathcal{F}_k) &\leq R(\widehat{f}_k) - (R(f_k^*) - \frac{\epsilon}{8}) \\
&= \max\{R_0(\widehat{f}_k), R_1(\widehat{f}_k)\} - \max\{R_0(f_k^*), R_1(f_k^*)\} + \frac{\epsilon}{8}. \qquad (19)
\end{aligned}$$

It follows by Theorem 1 that for m, n sufficiently large, we have

$$\begin{aligned}
\tilde{P}_0^m \otimes \tilde{P}_1^n(\sup_{f \in \mathcal{F}_k} |R_0(f) - \widehat{R}_0(f)| > \frac{\epsilon}{8}) &\leq \delta/2 \\
\tilde{P}_0^m \otimes \tilde{P}_1^n(\sup_{f \in \mathcal{F}_k} |R_1(f) - \widehat{R}_1(f)| > \frac{\epsilon}{8}) &\leq \delta/2.
\end{aligned}$$

Assume that both

$$|R_0(f) - \widehat{R}_0(f)| < \frac{\epsilon}{8} \qquad \text{for all } f \in \mathcal{F}_k$$
$$|R_1(f) - \widehat{R}_1(f)| < \frac{\epsilon}{8} \qquad \text{for all } f \in \mathcal{F}_k,$$

which by the result just stated, occurs with probability at least $1-\delta$ for m and n sufficiently large. It follows that

$$\max\{R_0(\widehat{f}_k), R_1(\widehat{f}_k)\} \quad < \quad \max\{\widehat{R}_0(\widehat{f}_k), \widehat{R}_1(\widehat{f}_k)\} + \frac{\epsilon}{8}$$

and

$$\max\{R_0(f_k^*), R_1(f_k^*)\} \quad > \quad \max\{\widehat{R}_0(f_k^*), \widehat{R}_1(f_k^*)\} - \frac{\epsilon}{8}.$$

Using these inequalities in Equation (19) yields

$$R(\widehat{f}_k) - R(\mathcal{F}_k) < \max\{\widehat{R}_0(\widehat{f}_k), \widehat{R}_1(\widehat{f}_k)\} + \frac{\epsilon}{8} - (\max\{\widehat{R}_0(f_k^*), \widehat{R}_1(f_k^*)\} - \frac{\epsilon}{8}) + \frac{\epsilon}{8}.$$

From our definition of $\widehat{f}_k$ in Equation (16), for m and n sufficiently large we have

$$\max\{\widehat{R}_0(\widehat{f}_k), \widehat{R}_1(\widehat{f}_k)\} \leq \max\{\widehat{R}_0(f_k^*), \widehat{R}_1(f_k^*)\} + \frac{\epsilon}{8}.$$

Therefore, we can conclude that

$$R(\widehat{f}_k) - R(\mathcal{F}_k) < \frac{\epsilon}{2},$$

with probability at least $1-\delta$. Thus, we conclude that

$$\tilde{P}_0^m \otimes \tilde{P}_1^n(R(\widehat{f}_k) - R^* < \epsilon) > 1 - \delta,$$

for m and n sufficiently large. ■

References

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.

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.

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

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

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

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.

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.

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

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.

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.

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.

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

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.

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

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

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.

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.

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

N. Manwani and P. S. Sastry. Noise tolerance under risk minimization. Technical Report arXiv:1109.5231, 2011.

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.

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

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

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

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.

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.