A Rate of Convergence for Mixture Proportion Estimation, with Application to Learning from Noisy Labels

Clayton Scott
University of Michigan, Department of EECS

Abstract

Mixture proportion estimation (MPE) is a fundamental tool for solving a number of weakly supervised learning problems – supervised learning problems where label information is noisy or missing. Previous work on MPE has established a universally consistent estimator. In this work we establish a rate of convergence for mixture proportion estimation under an appropriate distributional assumption, and argue that this rate of convergence is useful for analyzing weakly supervised learning algorithms that build on MPE. To illustrate this idea, we examine an algorithm for classification in the presence of noisy labels based on surrogate risk minimization, and show that the rate of convergence for MPE enables proof of the algorithm's consistency. Finally, we provide a practical implementation of mixture proportion estimation and demonstrate its efficacy in classification with noisy labels.

1 Introduction

Mixture proportion estimation (MPE) is the following problem: Let F, G, and H be probability distributions such that

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

where $0 \leq \kappa \leq 1$. Given random samples from F and H, estimate κ. It has recently been shown that a solution to MPE leads to solutions to various "weakly" supervised learning problems such as anomaly detection, learning from positive and unlabeled examples, domain adaptation, and classification with label noise (Blanchard et al., 2010; Scott et al., 2013; Sanderson and Scott, 2014).

Without any assumptions on F, G, and H, κ in (1) is not identifiable given F and H. 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. To ensure identifiability, it has been assumed that G is *irreducible* with respect to H, meaning that it is not possible to write $G = \gamma H + (1-\gamma)F'$, where F' is a distribution and $0 < \gamma \leq 1$. Blanchard et al. (2010) establish the following result.

Proposition 1 (Blanchard et al. (2010))**.** *If* (1) *holds and G is irreducible with respect to H, then κ in* (1) *is equal to*

$$\kappa^*(F|H) := \sup\{\kappa \,|\, F = (1-\kappa)G' + \kappa H \quad \textit{for some distribution } G'\},$$

the maximum proportion of H in F.

Blanchard et al. (2010) also establish a universally consistent estimator $\widehat{\kappa}$ for $\kappa^*(F|H)$, given random samples from F and H with sample sizes growing to infinity. By Proposition 1, this estimator also consistently estimates κ in (1) under the irreducibility assumption.

The first contribution of the present paper concerns the rate of convergence of $\widehat{\kappa}$ to κ^*. Blanchard et al. (2010) establish a "no free lunch" result which says that the rate of convergence of $\widehat{\kappa}$ to κ^* can be arbitrarily slow. In other words, to ensure a rate of convergence, it is necessary to make some kind of distributional assumption. We introduce such an assumption that is slightly stronger than irreducibility and ensures root-n rate of convergence for $\widehat{\kappa}$.

Our second contribution concerns the application of $\widehat{\kappa}$ to solve other learning problems like those mentioned above. We build on the recent work of Natarajan et al. (2013) who show that it is possible to learn a classifier in the presence of label noise by performing empirical risk minimization based on a cost-sensitive surrogate loss, where the cost parameter α of the loss is defined

Appearing in Proceedings of the 18^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2015, San Diego, CA, USA. JMLR: W&CP volume 38.

in terms of the label noise flipping probabilities. In Natarajan et al. (2013), these proportions are assumed known. When the proportions are *unknown*, which is more likely in practice, a consistent estimator $\widehat{\alpha}$ of α can be expressed in terms of $\widehat{\kappa}$, using the approach of Scott et al. (2013). Based on our first contribution, we show that a rate for $\widehat{\alpha}$, which follows from the rate for $\widehat{\kappa}$, leads to consistency of the surrogate-based learning procedure when the label noise proportions are unknown.

As a final contribution, we suggest a practical implementation of the estimator $\widehat{\kappa}$ of Blanchard et al. (2010), and demonstrate its efficacy on three data sets in a label noise setting, including a real data set from nuclear particle classification that is naturally described by the label noise model. Since $\widehat{\kappa}$ is based on Vapnik-Chervonenkis (VC) bounds, which are known to be loose for typical learning scenarios, its practical utility had not previously been clear.

An outline of the paper follows. The next section further motivates the MPE problem. Section 3 reviews the estimator $\widehat{\kappa}$ of Blanchard et al. (2010) and establishes a rate of convergence for this estimator under a proposed distributional assumption. Section 4 examines classification with surrogate losses under label noise with unknown noise proportions, and uses the results of Section 3 to establish a consistent classification procedure. In Section 5, $\widehat{\kappa}$ is demonstrated empirically and shown to be practically useful.

2 MPE for Weakly Supervised Learning

MPE is useful for a number of "weakly" supervised learning problems (WSL), wherein label information is missing or noisy in some way. For example, in the problem of learning from positive and unlabeled examples (LPUE) (Steinberg and Cardell, 1992; Denis, 1998; Liu et al., 2002; Denis et al., 2005; Elkan and Noto, 2008; Ward et al., 2009), H and G are the distributions governing the positive and negative classes, respectively, and F is the distribution of unlabeled examples. As has been previously observed, and as we show below, training a cost-sensitive classifier on the positive and unlabeled data yields a good cost-insensitive (i.e., conventional) classifier for the underlying problem of interest. However, the appropriate cost parameter depends on the proportion κ of positive examples in the unlabeled data, and since this is typically unknown, it needs to be estimated. Assuming G is irreducible with respect to H allows the estimator $\widehat{\kappa}$ of Blanchard et al. (2010) to be applied toward this end.

Another relevant WSL problem is binary classification with label noise, when the label noise is assumed to be independent of the observed feature vector (Blum and Mitchell, 1998; Lawrence and Schölkopf, 2001; Bouveyron and Girard, 2009; Stempfel and Ralaivola, 2009; Long and Servido, 2010; Manwani and Sastry, 2011; Natarajan et al., 2013). As with LPUE, training a cost-sensitive classifier on the noisy data yields a good cost-insensitive classifier. Once again, however, the appropriate cost parameter depends on the noise proportions. Scott et al. (2013) show that if each class-conditional distribution is irreducible with respect to the other, these proportions can be recovered by solving two MPE problems (see Proposition 3 below).

A third WSL problem that illustrates the utility of MPE is the following domain adaptation problem: given labeled examples from multiple classes, and an unlabeled data set where the classes occur in different proportions than in the training data, determine the proportion of each class in the unlabeled data set (Titterington, 1983; Latinne et al., 2001). Here F is the unlabeled data set, and H is any class-conditional distribution whose proportion in F is to be estimated. Sanderson and Scott (2014) demonstrate irreducibility-based assumptions under which MPE can be used to consistently estimate the unknown class proportions.

MPE has the potential to be applied to a number of other WSL problems, many of which reduce to one of the above problems at least in special cases. These include crowdsourcing (Raykar et al., 2010), multiple instance learning (Blum and Kalai, 1998), co-training (Blum and Mitchell, 1998), and learning from partial labels (Cour et al., 2011). Finally, we remark that MPE had been studied prior to Blanchard et al. (2010), but under parametric modeling assumptions (McLachlan, 1992; Bouveyron and Girard, 2009). Our focus is on nonparametric methods and learning-theoretic analysis.

3 A Rate for Estimation of the Maximal Mixing Proportion

To begin, we review the universally consistent estimator $\widehat{\kappa}$ of Blanchard et al. (2010) of $\kappa^*(F|H)$, which can converge arbitrarily slowly, and then introduce a distributional assumption under which this estimator converges at a known rate.

Let F and H be probability measures on a probability space $(\mathcal{X}, \mathfrak{S})$. Intuition for the estimator comes from a connection to the receiver operating characteristic (ROC) for the problem of testing the null hypothesis $X \sim H$ against the alternative $X \sim F$. Given a mea-

surable 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. (2010); see also Scott (2014).

Proposition 2 (Blanchard et al. (2010))**.**

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

In words, κ^* is the minimum slope among lines passing through the point $(1,1)$ in ROC space and some 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.

The estimator $\widehat{\kappa}$ of Blanchard et al. (2010) relies on VC theory (Devroye et al., 1996). Consider a sequence of VC classes of sets, $\mathcal{S}_1, \mathcal{S}_2, \ldots$, with VC dimensions $V_k < \infty$. 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 $i = 0, 1$, $\delta_i > 0$, $k \geq 1$ 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

$$\forall S \in \mathcal{S}_k \quad \left| Q(S) - \widehat{Q}(S) \right| \leq \epsilon_i(k, \delta_i), \tag{2}$$

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

In MPE we have training data

$$X_0^1, \ldots, X_0^{n_0} \overset{iid}{\sim} H, \tag{3}$$

$$X_1^1, \ldots, X_1^{n_1} \overset{iid}{\sim} F. \tag{4}$$

For $k \geq 1$, define

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

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 2, $\widehat{\kappa}(k, \delta_0, \delta_1)$ 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. (2010) show the following, which makes no assumption on the distributions F and H and thus establishes a universally consistent method for MPE.

Theorem 1 (Blanchard et al. (2010))**.** *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* $\min\{n_0, n_1\} \to \infty$.

We now introduce an assumption on F and H that will ensure a certain rate of convergence for $\widehat{\kappa}$ above. We use $\text{supp}(Q)$ to denote the support of a distribution Q.

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

The condition $\text{supp}(H) \not\subset \text{supp}(G)$ clearly implies that G is irreducible with respect to H, and therefore γ in **(A)** 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 $\text{supp}(H) \not\subset \text{supp}(G)$ 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 increasing 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, **(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 2. *Suppose* $(\mathcal{S}_k)_{k \geq 1}$ *is chosen to satisfy* **(AP2)**. *If* F *and* H *are such that* **(A)** *holds, then there exist a constant* $C > 0$ *such that for* n_0 *and* n_1 *sufficiently large, the estimator* $\widehat{\kappa}$ *satisfies*

$$\Pr\left(|\widehat{\kappa} - \kappa^*| \geq C \left[\sqrt{\frac{\log n_0}{n_0}} + \sqrt{\frac{\log n_1}{n_1}} \right] \right) \leq \frac{2}{n_0} + \frac{2}{n_1}. \tag{5}$$

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

Proof. We begin by establishing (5) without the absolute value, which is the more challenging direction. The reverse direction will follow easily by the first part of Theorem 1.

By **(A)**, there exists a distribution G and $\gamma \in [0,1]$ such that $F = (1-\gamma)G + \gamma H$ and $\text{supp}(H) \not\subset \text{supp}(G)$. Then G is irreducible with respect to H, and Proposition 1 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 1, 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. □

Henceforth we assume $\widehat{\kappa}$ is defined in terms of VC classes satisfying **(AP2)**.

4 Classification with Unknown Label Noise Proportions

We now study the use of surrogate losses for designing a classifier in the presence of label noise when the label noise proportions are unknown. We propose a natural learning rule for this problem and apply the rate of convergence result for $\widehat{\kappa}$ to deduce consistency of the learning procedure. Before addressing this problem, it is necessary to first review surrogate losses, and how they can be used to overcome label noise when the noise proportions are *known*.

Let (X, Y) be random on $\mathcal{X} \times \{0, 1\}$ where $\mathcal{X}$ is a measurable 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 (Devroye et al., 1996) that for any $f \in \mathcal{M}$, the excess P-risk satisfies

$$R_P(f) - R_P^* = 2\mathbb{E}_X[\mathbf{1}_{\{u(f(X)) \neq u(\eta(X) - \frac{1}{2})\}}|\eta(X) - \tfrac{1}{2}|], \tag{6}$$

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 $f \in \mathcal{M}$,

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

The corresponding Bayes risk is $R_{P,\alpha}^* := \inf_{f\in\mathcal{M}} R_{P,\alpha}(f)$, and the analogue to (6) is

$$R_\alpha(f) - R_\alpha^* = \mathbb{E}_X[\mathbf{1}_{\{u(f(X)) \neq u(\eta(X)-\alpha)\}}|\eta(X) - \alpha|] \tag{7}$$

(Scott, 2012). Note (6) corresponds to the case $\alpha = \frac{1}{2}$.

With this background, we now 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. 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. (2013) 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. Let us denote $\pi_i = \Pr(Y = 1 - i \,|\, \tilde{Y} = i)$, $i = 0, 1$, and introduce the following assumption on the amount of label noise.

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

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

Lemma 1. *If* **(B)** *holds, then for any* $f \in \mathcal{M}$,

$$R_P(f) - R_P^* = 2(1 - \pi_1 - \pi_0)(R_{\tilde{P},\alpha}(f) - R_{\tilde{P},\alpha}^*) \quad (8)$$

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

Proof. Note that **(B)** 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

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

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 (6) and (7):

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

□

4.1 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 Scott (2012), 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 $L(y, t) = \mathbf{1}_{\{y=1\}}L_1(t) + \mathbf{1}_{\{y=0\}}L_0(t)$. Given a loss L and $\alpha \in (0, 1)$, define

$$L_\alpha(y, t) := (1 - \alpha)\mathbf{1}_{\{y=1\}}L_1(t) + \alpha\mathbf{1}_{\{y=0\}}L_0(t). \quad (9)$$

Scott (2012) 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. (9). In particular, Bartlett et al. (2006) establish that if $L(y, t) = \phi(yt)$ 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 1 leads to the following result.

Corollary 1. *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. (2013) consider the setting where π_0 and π_1 are known. Using the above result, they apply Rademacher complexity analysis to bound $R_{\tilde{P},L_\alpha}(\widehat{f}) - R_{\tilde{P},L_\alpha}^*$ for a classification strategy $\widehat{f}$ based on a surrogate loss L_α.

4.2 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 the framework of Scott et al. (2013). Thus, suppose we observe noisy data

$$X_0^1, \ldots, X_0^{n_0} \overset{iid}{\sim} \tilde{P}_0 := (1 - \pi_0)P_0 + \pi_0 P_1, \quad (10)$$

$$X_1^1, \ldots, X_1^{n_1} \overset{iid}{\sim} \tilde{P}_1 := (1 - \pi_1)P_1 + \pi_1 P_0. \quad (11)$$

where P_0 and P_1 are the marginal distributions of X given $Y = 0$ and 1, respectively. The first sample can be thought of as observed patterns with noisy label $\tilde{Y} = 0$, and similarly for the second sample. For simplicity, the sample sizes n_0 and n_1 are assumed to be

nonrandom. Scott et al. (2013) establish the following result. The distributions P_0 and P_1 are said to be *mutually irreducible* if P_0 is irreducible with respect to P_1 and vice versa.

Proposition 3 (Scott et al. (2013))**.** *Assume* $\pi_0 + \pi_1 < 1$. *If* $P_0 \neq P_1$, *then* $\tilde{P}_1 \neq \tilde{P}_0$, *and there exist unique* $0 \leq \tilde{\pi}_0, \tilde{\pi}_1 < 1$ *such that*

$$\tilde{P}_0 = (1 - \tilde{\pi}_0)P_0 + \tilde{\pi}_0\tilde{P}_1 \tag{12}$$

$$\tilde{P}_1 = (1 - \tilde{\pi}_1)P_1 + \tilde{\pi}_1\tilde{P}_0. \tag{13}$$

In particular,

$$\tilde{\pi}_0 = \frac{\pi_0}{1 - \pi_1} < 1 \qquad \textit{and} \qquad \tilde{\pi}_1 = \frac{\pi_1}{1 - \pi_0} < 1. \tag{14}$$

Furthermore, if P_0 *and* P_1 *are mutually irreducible, then* $\tilde{\pi}_0 = \kappa^*(\tilde{P}_0|\tilde{P}_1)$ *and* $\tilde{\pi}_1 = \kappa^*(\tilde{P}_1|\tilde{P}_0)$.

Under the assumptions of this result, we can obtain estimates $\widehat{\tilde{\pi}}_0$ and $\widehat{\tilde{\pi}}_1$ of $\tilde{\pi}_0$ and $\tilde{\pi}_1$ using the estimator described in Section 3, and use these to estimate π_0 and π_1 by inverting the identities in (14), leading to the estimates

$$\widehat{\pi}_0 = \frac{\widehat{\tilde{\pi}}_0(1 - \widehat{\tilde{\pi}}_1)}{1 - \widehat{\tilde{\pi}}_0\widehat{\tilde{\pi}}_1} \qquad \text{and} \qquad \widehat{\pi}_1 = \frac{\widehat{\tilde{\pi}}_1(1 - \widehat{\tilde{\pi}}_0)}{1 - \widehat{\tilde{\pi}}_0\widehat{\tilde{\pi}}_1}. \tag{15}$$

To obtain a rate of convergence on $\widehat{\alpha}$, we need to ensure that $\tilde{P}_0$ and $\tilde{P}_1$ satisfy assumption **(A)** in both directions, and that P_0 and P_1 are mutually irreducible. The following assumption 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 4. *If* **(B)** *and* **(C)** *hold, then there exists* $C > 0$ *such that for* n_0 *and* n_1 *sufficiently large,*

$$\Pr\left(|\widehat{\alpha} - \alpha| \geq C\left[\sqrt{\frac{\log n_0}{n_0}} + \sqrt{\frac{\log n_1}{n_1}}\right]\right) \leq \frac{4}{n_0} + \frac{4}{n_1}.$$

Proof. **(B)** implies $\pi_0 + \pi_1 < 1$, and by **(C)**, P_0 and P_1 are mutually irreducible which further implies $P_0 \neq P_1$. Thus Proposition 3 implies $\tilde{\pi}_0 = \kappa^*(\tilde{P}_0|\tilde{P}_1)$ and $\tilde{\pi}_1 = \kappa^*(\tilde{P}_1|\tilde{P}_0)$. Next, apply Theorem 2 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 **(A)** for both $(F, H) = (\tilde{P}_1, \tilde{P}_0)$ and $(F, H) = (\tilde{P}_0, \tilde{P}_1)$. We will show **(A)** for $(F, H) = (\tilde{P}_1, \tilde{P}_0)$, the other case being similar. From (13), it suffices to show $\text{supp}(\tilde{P}_1) \not\subset \text{supp}(P_0)$. But this holds because $\tilde{P}_1 = (1 - \pi_1)P_1 + \pi_1 P_0$ (see Eqn. (11)) and $\text{supp}(P_1) \not\subset \text{supp}(P_0)$ and $\pi_1 < 1$. We can now apply Theorem 2 to both $\widehat{\tilde{\pi}}_0$ and $\widehat{\tilde{\pi}}_1$. It is then not hard to show that these rates lead to similar rates for $\widehat{\pi}_1$ and $\widehat{\pi}_0$, which in turn lead to the desired rate for $\widehat{\alpha}$. □

4.3 Algorithm and Main Consistency Result

We now introduce a consistent classification procedure based on surrogate losses in the case of unknown label noise proportions. In addition to the two data sets used to estimate α, we assume a third data set

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

where $\tilde{P}$ is the joint distribution of $(X, \tilde{Y})$. For simplicity, we assume that n, n_0, and n_1 are of the same order when tending to infinity. **Remark:** Although we are assuming separating training sets for the classifier and for $\widehat{\alpha}$, this is actually not necessary. Having two training sets just makes the analysis slightly simpler since n_0 and n_1 are nonrandom in our setup.

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

$$\widehat{f} = \arg\min_{f \in \mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} L_{\widehat{\alpha}}(\tilde{Y}_i, f(X_i)) + \lambda_n \|f\|_{\mathcal{H}}^2, \tag{16}$$

where $L_{\widehat{\alpha}}$ is the $\widehat{\alpha}$-weighted cost-sensitive loss associated with L, as defined in (9). For example, if $L(y, t) = \max\{0, 1 - yt\}$ is the hinge loss, $\widehat{f}$ is a cost-sensitive support vector machine.

We will assume that the reproducing kernel k associated with $\mathcal{H}$ is universal and bounded (Steinwart and Christmann, 2008). 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 3. *Assume* **(B)** *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*

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

as $n_0, n_1, n \to \infty$.

Proof. By Corollary 1, it suffices to show $R_{\tilde{P}, L_\alpha}(\widehat{f}) - R^*_{\tilde{P}, L_\alpha} \to 0$ in probability. 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 $J(f) := \widehat{R}_{L_{\widehat{\alpha}}}(f) + \lambda_n\|f\|^2$. Also define $L_0 := \max\{L(0,0), L(1,0)\}$. Observe that $J(\widehat{f}) \leq J(0) \leq L_0$. Therefore $\lambda_n\|\widehat{f}\|^2 \leq L_0 - \widehat{R}_{L_{\widehat{\alpha}}}(\widehat{f}) \leq L_0$, and we deduce that $\widehat{f} \in B_{\mathcal{H}}(M_n)$, the ball of radius M_n in $\mathcal{H}$, where $M_n := \sqrt{L_0/\lambda_n}$.

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 (Steinwart and Christmann, 2008). Then

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

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

$$\frac{2DBM_n}{\sqrt{n}} + 2BM_n\sqrt{\frac{\ln 2n}{2n}}$$

using Rademacher complexity analysis for balls in a RKHS (Mohri et al., 2012). Here D is the Lipschitz constant for L and B is the bound on the kernel. By the assumed rate of decay for λ_n, both term tend to zero as $n \to \infty$. For the last term, we also need to observe 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 $\widehat{f}$, since $J(\widehat{f}) \leq J(f_\epsilon)$ implies $\widehat{R}_{L_{\widehat{\alpha}}}(\widehat{f}) - \widehat{R}_{L_{\widehat{\alpha}}}(f_\epsilon) \leq \lambda_n\|f_\epsilon\|^2 - \lambda_n\|\widehat{f}\|^2 \leq \lambda_n\|f_\epsilon\|^2$.

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

$$\begin{aligned} \widehat{R}_{L_\alpha}(f) - \widehat{R}_{L_{\widehat{\alpha}}}(f) &= \frac{1}{n}\Bigg[\sum_{i:\tilde{Y}_i=1} (\widehat{\alpha} - \alpha)L(1, f(X_i)) \\ &\qquad + \sum_{i:\tilde{Y}_i=0} (\alpha - \widehat{\alpha})L(0, f(X_i))\Bigg] \\ &\leq |\widehat{\alpha} - \alpha| \sup_{x,y} L(y, f(x)) \\ &\leq |\widehat{\alpha} - \alpha| \left(L_0 + D\|f\|_\infty\right), \end{aligned}$$

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 $\|f\|_{\mathcal{H}} \leq \sqrt{\frac{L_0}{\lambda_n}}$, and so for the second term to go to zero, we need $|\widehat{\alpha} - \alpha|/\lambda_n$ to go to zero. Under **(B)** 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 $R_{\tilde{P},L_\alpha}(\widehat{f}) - R^*_{\tilde{P},L_\alpha} \leq \epsilon$ with probability tending to one as n (and with it n_0 and n_1) tends to infinity. This concludes the proof. □

5 Implementation and Experiments for Mixture Proportion Estimation

The estimator $\widehat{\kappa}$ relies on VC bounds, which are known to be loose in typical learning situations. Therefore it is not obvious that the estimator $\widehat{\kappa}$ is practically useful, not to mention tractable. In this section, we propose an implementation of the mixture proportion estimator that is closely motivated by $\widehat{\kappa}$, and demonstrate its performance on three data sets.

$\widehat{\kappa}$ works as follows: Consider the collection of classifiers $\mathcal{S}_1 \cup \mathcal{S}_2 \cup \cdots$. For each f in this collection, conservatively estimate the false positive and true positive probabilities, such that their ratio is an upper bound on κ^*. If the collection of classifiers is rich enough, and the sample sizes tend to infinity, this upper bound converges to κ^*.

Motivated by this idea, we suggest the following practical algorithm for MPE. First, split each of the two samples (10) and (11) in half (or some other ratio). Using the first half 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. Using the second half of each sample, construct conservative estimates 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, which is the tightest possible deviation bound for a binomial random variable (Langford, 2005). Using these conservative estimates, we then compute the minimum slope for any of these (conservatively estimated) operating points to the point $(1, 1)$ in ROC space. A Matlab implementation is available at `http://web.eecs.umich.edu/~cscott/code.html`.

To study the performance of this implementation, 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 (Adams and White, 1978). It is important to classify gamma-ray pulses from neutron pulses, because the energy distribution of neutrons is used to characterize different nuclear materials in nuclear inspection settings (e.g., to inventory nuclear materials at nuclear power facilities). Training data was obtained by measuring particles emitted from a Cf-252 source, which undergoes spontaneous decay and emits both neutrons and gamma rays. Through a special experimental configuration, 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 can be obtained by focusing on those particles with TOFs around the speed of light. 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. Similarly, a TOF-window to select neutrons will also contain some proportion of gamma-ray pulses. We obtain samples of size $n_0 = n_1 = 3000$ from the Cf-252 source. 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. These results indicate that our implementation provides reasonably accurate estimates of the label noise proportions in the four experimental settings where the true proportions are known. In the nuclear particle classification problem, although ground truth labels are unavailable, the estimated proportions are at least consistent with the expectation that noisy labels should be relatively rare (given the high rate of fission events relative to the expected rate of background events), and also with the knowledge that neutrons are rarer background events than gamma-rays.

data set	π_0	π_1	$\widehat{\pi}_0$	$\widehat{\pi}_1$
waveform	0.1	0.25	0.1072	0.2808
waveform	0.15	0.05	0.1470	0.0679
digits	0.1	0.25	0.1153	0.1955
digits	0.15	0.05	0.1432	0.0419
nuclear	N/A	N/A	0.0185	0.0812

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

6 Final Thoughts

We have demonstrated a distributional assumption for MPE under which the universally consistent estimator of Blanchard et al. (2010) converges at a known rate. We then applied this result to establish the consistency of a surrogate-based algorithm for classification with label noise with unknown noise proportions. Although we have focused on the risk as a performance measure, our rate of convergence result should also be instrumental in establishing consistency or rates of convergence for learning algorithms geared toward other performance measures such as the F-measure.

We also proposed an implementation of $\widehat{\kappa}$ and demonstrated reasonable performance in a label noise setup. It would also be natural to experimentally examine the performance of a classifier trained according to the algorithm in (16). Fortunately, this has already been done in a certain sense. Natarajan et al. (2013) examine the sensitivity of their algorithm, which assumes known noise proportions, to misspecification of these proportions. When the misspecification is minor, the decrease in performance is negligible, suggesting that accurate mixture proportion estimation will indeed translate to accurate classification.

Acknowledgements

The author thanks Marek Flaska and Tyler Sanderson for the nuclear particle data, and was supported in part by NSF Grants 0953135, 1047871, 1217880, and 1422157.

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 spec-

troscopy. *Nuclear Instruments and Methods in Physics Research*, 1978.

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

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.

Avrim Blum and Adam Kalai. A note on learning from multiple-instance examples. *Machine Learning*, 30 (1):23–29, 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.

T. Cour, B. Sapp, and B. Taskar. Learning from partial labels. *J. Machine Learning Research*, 12:1501–1536, 2011.

F. Denis. PAC learning from positive statistical queries. In *Proc. 9th Int. Conf. on Algorithmic Learning Theory (ALT)*, pages 112–126, Otzenhausen, Germany, 1998.

F. Denis, R. Gilleron, and F. Letouzey. Learning from positive and unlabeled examples. *Theoretical Computer Science*, 348(1):70–83, 2005.

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

C. Elkan and K. Noto. Learning classifiers from only positive and unlabeled data. In *Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD08)*, pages 213–220, 2008.

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

P. Latinne, M. Saerens, and C. Decaestecker. Adjusting the outputs of a classier to new a priori probabilities may signicantly improve classication accuracy: Evidence from a multi-class problem in remote sensing. In C. Sammut and A. H. Hoffmann, editors, *Proc. 18th Int. Conf. on Machine Learning*, pages 298–305, 2001.

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.

B. Liu, W. S. Lee, P. S. Yu, and X. Li. Partially supervised classification of text documents. In *Proc. 19th Int. Conf. Machine Learning (ICML)*, pages 387–394, Sydney, Australia, 2002.

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. *IEEE Trans. on Cybernetics*, 43 (3):1146–1151, 2011.

G. McLachlan. *Discriminant Analysis and Statistical Pattern Recognition*. Wiley, 1992.

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

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

V. C. Raykar, S. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bogoni, and L. Moy. Learning from crowds. *The Journal of Machine Learning Research*, 99:1297–1322, 2010.

T. Sanderson and C. Scott. Class proportion estimation with application to multiclass anomaly rejection. In *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2014.

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

C. Scott. Notes on weakly supervised learning, 2014. URL `web.eecs.umich.edu/∼cscott/wsl.pdf`.

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.

Dan Steinberg and N. Scott Cardell. Estimating logistic regression models when the dependent variable has no variance. *Communications in Statistics - Theory and Methods*, 21:423–450, 1992.

I. Steinwart and A. Christmann. *Support Vector Machines*. Springer, 2008.

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.

D. M. Titterington. Minimum distance nonparametric estimation of mixture proportions. *Journal of the Royal Statistical Society*, 45(1):37–46, 1983.

G. Ward, T. Hastie, S. Barry, J. Elith, and J. R. Leathwick. Presence-only data and the EM algorithm. *Biometrics*, 65:554–564, 2009.