Mixture Proportion Estimation via Kernel Embedding of Distributions

Harish G. Ramaswamy[†] HGRAMASW@IN.IBM.COM
IBM Research, Bangalore, India
Indian Institute of Science, Bangalore, India
Clayton Scott CLAYSCOT@UMICH.EDU
EECS and Statistics, University of Michigan, Ann Arbor, USA
Ambuj Tewari TEWARIA@UMICH.EDU
Statistics and EECS, University of Michigan, Ann Arbor, USA

Abstract

Mixture proportion estimation (MPE) is the problem of estimating the weight of a component distribution in a mixture, given samples from the mixture and component. This problem constitutes a key part in many "weakly supervised learning" problems like learning with positive and unlabelled samples, learning with label noise, anomaly detection and crowdsourcing. While there have been several methods proposed to solve this problem, to the best of our knowledge no efficient algorithm with a proven convergence rate towards the true proportion exists for this problem. We fill this gap by constructing a provably correct algorithm for MPE, and derive convergence rates under certain assumptions on the distribution. Our method is based on embedding distributions onto an RKHS, and implementing it only requires solving a simple convex quadratic programming problem a few times. We run our algorithm on several standard classification datasets, and demonstrate that it performs comparably to or better than other algorithms on most datasets.

1. Introduction

Mixture proportion estimation (MPE) is the problem of estimating the weight of a component distribution in a mixture, given samples from the mixture and component. Solving this problem happens to be a key step in solving several "weakly supervised" learning problems. For example, MPE is a crucial ingredient in solving the weakly supervised learning problem of learning from positive and unlabelled samples (LPUE), in which one has access to unlabelled data and positively labelled data but wishes to construct a classifier distinguishing between positive and negative data (Liu et al., 2002; Denis et al., 2005; Ward et al., 2009). MPE also arises naturally in the task of learning a classifier with noisy labels in the training set, i.e., positive instances have a certain chance of being mislabelled as negative and vice-versa, independent of the observed feature vector (Lawrence & Scholkopf, 2001; Bouveyron & Girard, 2009; Stempfel & Ralaivola, 2009; Long & Servido, 2010; Natarajan et al., 2013). Natarajan et al. (2013) show that this problem can be solved by minimizing an appropriate cost sensitive loss. But the cost parameter depends on the label noise parameters, the computation of which can be broken into two MPE problems (Scott et al., 2013a). MPE also has applications in several other problems like anomaly rejection (Sanderson & Scott, 2014) and crowdsourcing (Raykar et al., 2010).

When no assumptions are made on the mixture and the components, the problem is ill defined as the mixture proportion is not identifiable (Scott, 2015). While several methods have been proposed to solve the MPE problem (Blanchard et al., 2010; Sanderson & Scott, 2014; Scott, 2015; Elkan & Noto, 2008; du Plessis & Sugiyama, 2014; Jain et al., 2016), to the best of our knowledge no provable and efficient method is known for solving this problem in the general non-parametric setting with minimal assumptions. Some papers propose estimators that converge to the true proportion under certain conditions (Blanchard et al., 2010; Scott et al., 2013a; Scott, 2015), but they cannot be efficiently computed. Hence they use a method which is motivated based on the provable method but has no direct guarantees of convergence to the true proportion. Some papers propose an estimator that can be implemented efficiently (Elkan & Noto, 2008; du Plessis & Sugiyama,

[†]Part of work done when visiting the University of Michigan.
Proceedings of the 33^{rd} International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48.

2014), but the resulting estimator is correct only under very restrictive conditions (see Section 7) on the distribution. Further, all these methods except the one by du Plessis & Sugiyama (2014) require an accurate binary conditional probability estimator as a sub-routine and use methods like logistic regression to achieve this. In our opinion, requiring an accurate conditional probability estimate (which is a real valued function over the instance space) for estimating the mixture proportion (a single number) is too roundabout.

Our main contribution in this paper is an efficient algorithm for mixture proportion estimation along with convergence rates of the estimate to the true proportion (under certain conditions on the distribution). The algorithm is based on embedding the distributions (Gretton et al., 2012) into a reproducing kernel Hilbert space (RKHS), and only requires a simple quadratic programming solver as a sub-routine. Our method does not require the computation of a conditional probability estimate and is hence potentially better than other methods in terms of accuracy and efficiency. We test our method on some standard datasets, compare our results against several other algorithms designed for mixture proportion estimation and find that our method performs better than or comparable to previously known algorithms on most datasets.

The rest of the paper is organised as follows. The problem set up and notations are given in Section 2. In Section 3 we introduce the main object of our study, called the $\mathcal{C}$-distance, which essentially maps a candidate mixture proportion value to a measure of how 'bad' the candidate is. We give a new condition on the mixture and component distributions that we call 'separability' in Section 4, under which the $\mathcal{C}$-distance function explicitly reveals the true mixture proportion, and propose two estimators based on this. In Section 5 we give the rates of convergence of the proposed estimators to the true mixture proportion. We give an explicit implementation of one of the estimators based on a simple binary search procedure in Section 6. We give brief summaries of other known algorithms for mixture proportion estimation in Section 7 and list their characteristics and shortcomings. We give details of our experiments in Section 8 and conclude in Section 9.

2. Problem Setup and Notations

Let G, H be distributions over a compact metric space $\mathcal{X}$ with supports given by $\text{supp}(G), \text{supp}(H)$. Let $\kappa^* \in [0, 1)$ and let F be a distribution that is given by a convex combination (or equivalently, a mixture) of G and H as follows:

$$F = (1 - \kappa^*)G + \kappa^* H.$$

Equivalently, we can write

$$G = (\lambda^*)F + (1 - \lambda^*)H,$$

where $\lambda^* = \frac{1}{1-\kappa^*}$. Given samples $\{x_1, x_2, \ldots, x_n\}$ drawn i.i.d. from F and $\{x_{n+1}, \ldots, x_{n+m}\}$ drawn i.i.d. from H, the objective in mixture proportion estimation (MPE) (Scott, 2015) is to estimate κ^*.

Let $\mathcal{H}$ be a reproducing kernel Hilbert space (RKHS) (Aronszajn, 1950; Berlinet & Thomas, 2004) with a positive semi-definite kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Let $\phi : \mathcal{X} \to \mathcal{H}$ represent the kernel mapping $x \mapsto k(x, .)$. For any distribution P over $\mathcal{X}$, let $\phi(P) = \mathbf{E}_{X\sim P}\phi(X)$. It can be seen that for any distribution P and $f \in \mathcal{H}$, that $\mathbf{E}_{X\sim P} f(X) = \langle f, \phi(P)\rangle_{\mathcal{H}}$. Let $\Delta_{n+m} \subseteq \mathbb{R}^{n+m}$ be the $(n + m - 1)$-dimensional probability simplex given by $\Delta_{n+m} = \{\mathbf{p} \in [0, 1]^{n+m} : \sum_i p_i = 1\}$. Let $\mathcal{C}, \mathcal{C}_S$ be defined as

$$\mathcal{C} = \{w \in \mathcal{H} : w = \phi(P), \text{for some distribution } P\},$$

$$\mathcal{C}_S = \{w \in \mathcal{H} : w = \sum_{i=1}^{n+m} \alpha_i \phi(x_i), \text{ for some } \alpha \in \Delta_{n+m}\}.$$

Clearly, $\mathcal{C}_S \subseteq \mathcal{C}$, and both $\mathcal{C}, \mathcal{C}_S$ are convex sets.

Let $\widehat{F}$ be the distribution over $\mathcal{X}$ that is uniform over $\{x_1, x_2, \ldots, x_n\}$. Let $\widehat{H}$ be the distribution over $\mathcal{X}$ that is uniform over $\{x_{n+1}, \ldots, x_{n+m}\}$. As F is a mixture of G and H, we have that some $S_1 \subseteq \{x_1, \ldots, x_n\}$ is drawn from G and the rest from H. We let $\widehat{G}$ denote the uniform distribution over S_1. On average, we expect the cardinality of S_1 to be $\frac{n}{\lambda^*}$. Note that we do not know S_1 and hence cannot compute $\phi(\widehat{G})$ directly, however we have that $\phi(\widehat{G}) \in \mathcal{C}_S$.

3. RKHS Distance to Valid Distributions

Define the "$\mathcal{C}$-distance" function $d : [0, \infty) \to [0, \infty)$ as follows:

$$d(\lambda) = \inf_{w\in\mathcal{C}} \|\lambda\phi(F) + (1 - \lambda)\phi(H) - w\|_{\mathcal{H}}. \quad (1)$$

Intuitively, $d(\lambda)$ reconstructs $\phi(G)$ from F and H assuming $\lambda^* = \lambda$, and computes its distance to $\mathcal{C}$. Also, define the empirical version of the $\mathcal{C}$-distance function, $\widehat{d} : [0, \infty) \to [0, \infty)$, which we call the $\mathcal{C}_S$-distance function, as

$$\widehat{d}(\lambda) = \inf_{w\in\mathcal{C}_S} \|\lambda\phi(\widehat{F}) + (1 - \lambda)\phi(\widehat{H}) - w\|_{\mathcal{H}}\,. \quad (2)$$

Note that the $\mathcal{C}_S$-distance function $\widehat{d}(\lambda)$ can be computed efficiently via solving a quadratic program. For any $\lambda \geq 0$, let $\mathbf{u}_\lambda \in \mathbb{R}^{n+m}$ be such that $\mathbf{u}_\lambda^\top = \frac{\lambda}{n}([\mathbf{1}_n^\top, \mathbf{0}_m^\top]) + \frac{1-\lambda}{m}([\mathbf{0}_n^\top, \mathbf{1}_m^\top])$, where $\mathbf{1}_n$ is the n-dimensional all ones vector, and $\mathbf{0}_m$ is the m-dimensional all zeros vector. Let $K \in \mathbb{R}^{(n+m)\times(n+m)}$ be the kernel matrix given by $K_{i,j} = k(x_i, x_j)$. We then have

$$(\widehat{d}(\lambda))^2 = \inf_{\mathbf{v}\in\Delta_{n+m}} (\mathbf{u}_\lambda - \mathbf{v})^\top K(\mathbf{u}_\lambda - \mathbf{v})\,.$$

We now give some basic properties of the $\mathcal{C}$-distance function and the $\mathcal{C}_S$-distance function that will be of use later. All proofs not found in the paper can be found in the supplementary material.

Proposition 1.

$$d(\lambda) = 0, \qquad \forall \lambda \in [0, \lambda^*],$$
$$\widehat{d}(\lambda) = 0, \qquad \forall \lambda \in [0, 1]\,.$$

Proposition 2. *$d(.)$ and $\widehat{d}(.)$ are non-decreasing convex functions on $[0, \infty)$.*

Below, we give a simple reformulation of the $\mathcal{C}$-distance function and basic lower and upper bounds that reveal its structure.

Proposition 3. *For all $\mu \geq 0$,*

$$d(\lambda^* + \mu) = \inf_{w \in \mathcal{C}} \|\phi(G) + \mu(\phi(F) - \phi(H)) - w\|_{\mathcal{H}}\,.$$

Proposition 4. *For all $\lambda, \mu \geq 0$,*

$$d(\lambda) \geq \lambda \|\phi(F) - \phi(H)\| - \sup_{w \in \mathcal{C}} \|\phi(H) - w\|, \quad (3)$$
$$d(\lambda^* + \mu) \leq \mu \|\phi(F) - \phi(H)\|\,. \quad (4)$$

Using standard results of Smola et al. (2007), we can show that the kernel mean embeddings of the empirical versions of F, H and G are close to the embeddings of the distributions themselves.

Lemma 5. *Let the kernel k be such that $k(x, x) \leq 1$ for all $x \in \mathcal{X}$. Let $\delta \in (0, 1/4]$. The following holds with probability $1 - 4\delta$ (over the sample $x_1, \ldots, x_{n+m}$) if $n > 2(\lambda^*)^2 \log\left(\frac{1}{\delta}\right)$,*

$$\|\phi(F) - \phi(\widehat{F})\|_{\mathcal{H}} \leq \frac{3\sqrt{\log(1/\delta)}}{\sqrt{n}}$$
$$\|\phi(H) - \phi(\widehat{H})\|_{\mathcal{H}} \leq \frac{3\sqrt{\log(1/\delta)}}{\sqrt{m}}$$
$$\|\phi(G) - \phi(\widehat{G})\|_{\mathcal{H}} \leq \frac{3\sqrt{\log(1/\delta)}}{\sqrt{n/(2\lambda^*)}}.$$

We will call this $1 - 4\delta$ high probability event as E_δ. All our results hold under this event.

Using Lemma 5 one can show that the $\mathcal{C}$-distance function and the $\mathcal{C}_S$-distance function are close to each other. Of particular use to us is an upper bound on the $\mathcal{C}_S$-distance function $\widehat{d}(\lambda)$ for $\lambda \in [1, \lambda^*]$, and a general lower bound on $\widehat{d}(\lambda) - d(\lambda)$.

Lemma 6. *Let $k(x, x) \leq 1$ for all $x \in \mathcal{X}$. Assume E_δ. For all $\lambda \in [1, \lambda^*]$ we have that*

$$\widehat{d}(\lambda) \leq \left(2 - \frac{1}{\lambda^*} + \frac{\sqrt{2}}{\sqrt{\lambda^*}}\right) \lambda \cdot \frac{3\sqrt{\log(1/\delta)}}{\sqrt{\min(m, n)}}.$$

Lemma 7. *Let $k(x, x) \leq 1$ for all $x \in \mathcal{X}$. Assume E_δ. For all $\lambda \geq 1$, we have*

$$\widehat{d}(\lambda) \geq d(\lambda) - (2\lambda - 1) \cdot \frac{3\sqrt{\log(1/\delta)}}{\sqrt{\min(m, n)}}\,.$$

4. Mixture Proportion Estimation under a Separability Condition

Blanchard et al. (2010); Scott (2015) observe that without any assumptions on F, G and H, the mixture proportion κ^* is not identifiable, and postulate an "irreducibility" assumption under which κ^* becomes identifiable. The irreducibility assumption essentially states that G cannot be expressed as a non-trivial mixture of H and some other distribution. Scott (2015) propose a stronger assumption than irreducibility under which they provide convergence rates of the estimator proposed by Blanchard et al. (2010) to the true mixture proportion κ^*. We call this condition as the "anchor set" condition as it is similar to the "anchor words" condition of Arora et al. (2012) when the domain $\mathcal{X}$ is finite.

Definition 8. *A family of subsets $\mathcal{S} \subseteq 2^{\mathcal{X}}$, and distributions G, H are said to satisfy the* anchor set condition with margin $\gamma > 0$, *if there exists a compact set $A \in \mathcal{S}$ such that $A \subseteq \text{supp}(H) \setminus \text{supp}(G)$ and $H(A) \geq \gamma$.*

We propose another condition which is similar to the anchor set condition (and is defined for a class of functions on $\mathcal{X}$ rather than subsets of $\mathcal{X}$). Under this condition we show that the $\mathcal{C}$-distance function (and hence the $\mathcal{C}_S$-distance function) reveals the true mixing proportion λ^*.

Definition 9. *A class of functions $\mathcal{H} \subseteq \mathbb{R}^{\mathcal{X}}$, and distributions G, H are said to satisfy* separability condition with margin $\alpha > 0$ and tolerance β, *if $\exists h \in \mathcal{H}, \|h\|_{\mathcal{H}} \leq 1$ and*

$$\mathbf{E}_{X \sim G} h(X) \leq \inf_x h(x) + \beta \leq \mathbf{E}_{X \sim H} h(X) - \alpha\,.$$

We say that a kernel k and distributions G, H satisfy the separability condition, if the unit norm ball in its RKHS and distributions G, H satisfy the separability condition.

Given a family of subsets satisfying the anchor set condition with margin γ, it can be easily seen that the family of functions given by the indicator functions of the family of subsets satisfy the separability condition with margin $\alpha = \gamma$ and tolerance $\beta = 0$. Hence this represents a natural extension of the anchor set condition to a function space setting.

Under separability one can show that λ^* is the "departure point from zero" for the $\mathcal{C}$-distance function.

Theorem 10. *Let the kernel k, and distributions G, H satisfy the separability condition with margin $\alpha > 0$ and tol-*

erance β. *Then* $\forall \mu > 0$

$$d(\lambda^* + \mu) \geq \frac{\alpha \mu}{\lambda^*} - \beta \,.$$

Proof. (Sketch) For any inner product $\langle ., . \rangle$ and its norm $\|.\|$ over the vector space $\mathcal{H}$, we have that $\|f\| \geq \langle f, g \rangle$ for all $g \in \mathcal{H}$ with $\|g\| \leq 1$. The proof mainly follows by lower bounding the norm in the definition of $d(.)$, with an inner product with the witness g of the separability condition. □

Further, one can link the separability condition and the anchor set condition via universal kernels (like the Gaussian RBF kernel) (Michelli et al., 2006), which are kernels whose RKHS is dense in the space of all continuous functions over a compact domain.

Theorem 11. *Let the kernel* $k : \mathcal{X} \times \mathcal{X} \to [0, \infty)$ *be universal. Let the distributions* G, H *be such that they satisfy the anchor set condition with margin* $\gamma > 0$ *for some family of subsets of* $\mathcal{X}$. *Then, for all* $\theta > 0$, *there exists a* $\beta > 0$ *such that the kernel* k, *and distributions* G, H *satisfy the separability condition with margin* $\beta\theta$ *and tolerance* β.

Proof. (Sketch) As the distributions G, H satisfy the anchor set condition, there must exist a continuous non-negative function that is zero on the support of G and greater than one on the set A that witnesses the anchor set condition. Due to universality of the kernel k, there must exist an element in its RKHS that arbitrarily approximates this function. The normalised version of this function forms a witness to the separability condition. □

The ultimate objective in mixture proportion estimation is to estimate κ^* (or equivalently λ^*). If one has direct access to $d(.)$ and the kernel k and distributions G, H satisfy the separability condition with tolerance $\beta = 0$, then we have by Proposition 1 and Theorem 10 that

$$\lambda^* = \inf\{\lambda : d(\lambda) > 0\}.$$

We do not have direct access to $d(.)$, but we can calculate $\widehat{d}(.)$. From Lemmas 1 and 7, we have that for all $\lambda \in [0, \lambda^*]$, $\widehat{d}(\lambda)$ converges to 0 as the sample size $\min(m, n)$ increases. From Lemma 7 we have that for all $\lambda \geq 0$, $\widehat{d}(\lambda) \geq d(\lambda) - \epsilon$ for any $\epsilon > 0$ if $\min(m, n)$ is large enough. Hence $\widehat{d}(.)$ is a good surrogate for $d(.)$ and based on this observation we propose two strategies of estimating λ^* and show that the errors of both these strategies can be made to approach 0 under the separability condition.

The first estimator is called the value thresholding estimator. For some $\tau \in [0, \infty)$ it is defined as,

$$\widehat{\lambda}^V_\tau = \inf\{\lambda : \widehat{d}(\lambda) \geq \tau\} \,.$$

The second estimator is called the gradient thresholding estimator. For some $\nu \in [0, \infty)$ it is defined as

$$\widehat{\lambda}^G_\nu = \inf\{\lambda : \exists g \in \partial \widehat{d}(\lambda), g \geq \nu\},$$

where $\partial \widehat{d}(\lambda)$ is the sub-differential of $\widehat{d}(.)$ at λ. As $\widehat{d}(.)$ is a convex function, the slope of $\widehat{d}(.)$ is a non-decreasing function and thus thresholding the gradient is also a viable strategy for estimating λ^*.

To illustrate some of the ideas above, we plot $\widehat{d}(.)$ and $\nabla \widehat{d}(.)$ for two different true mixing proportions κ^* and sample sizes in Figure 2. The data points from the component and mixture distribution used for computing the plot are taken from the `waveform` dataset.

5. Convergence of Value and Gradient Thresholding Estimators

We now show that both the value thresholding estimator $\widehat{\lambda}^V_\tau$ and the gradient thresholding estimator $\widehat{\lambda}^G_\nu$ converge to λ^* under appropriate conditions.

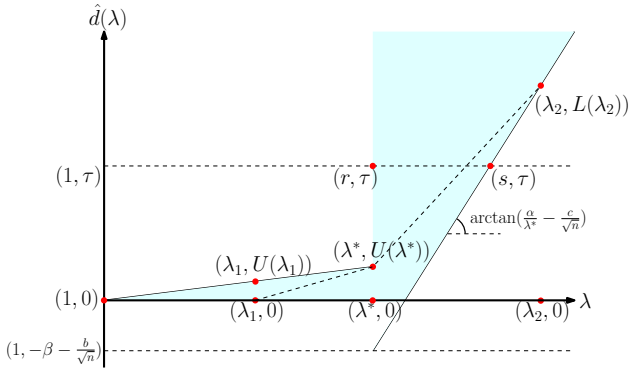
Theorem 12. *Let* $\delta \in (0, \frac{1}{4}]$. *Let* $k(x, x) \leq 1$ *for all* $x \in \mathcal{X}$. *Let the kernel* k, *and distributions* G, H *satisfy the separability condition with tolerance* β *and margin* $\alpha > 0$. *Let the number of samples be large enough such that* $\min(m, n) > \frac{(12 \cdot \lambda^*)^2 \log(1/\delta)}{\alpha^2}$. *Let the threshold* τ *be such that* $\frac{3\lambda^* \sqrt{\log(1/\delta)}(2 - 1/\lambda^* + \sqrt{2/\lambda^*})}{\sqrt{\min(m,n)}} \leq \tau \leq \frac{6\lambda^* \sqrt{\log(1/\delta)}(2 - 1/\lambda^* + \sqrt{2/\lambda^*})}{\sqrt{\min(m,n)}}$. *We then have with probability* $1 - 4\delta$

$$\lambda^* - \widehat{\lambda}^V_\tau \leq 0,$$
$$\widehat{\lambda}^V_\tau - \lambda^* \leq \frac{\beta \lambda^*}{\alpha} + c \cdot \sqrt{\log(1/\delta)} \cdot (\min(m, n))^{-1/2},$$

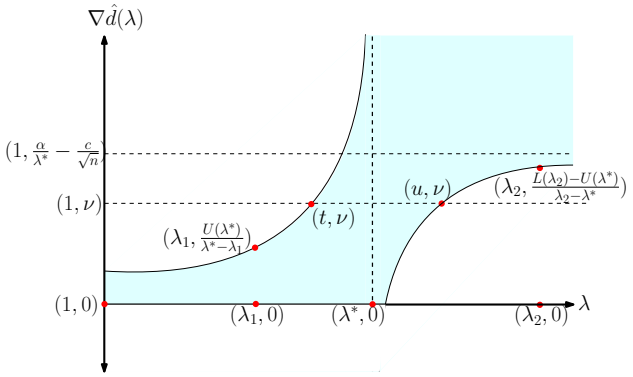
where $c = \left(\frac{6\alpha(\lambda^*)^2(2 - 1/\lambda^* + \sqrt{2/\lambda^*}) + 2\lambda^*(3\alpha + 6\lambda^*(2 + \alpha + \beta))}{\alpha^2} \right)$.

Proof. (Sketch) Under event E_δ, Lemma 6 gives an upper bound on $\widehat{d}(\lambda)$ for $\lambda \in [1, \lambda^*]$, which is denoted by the line $(\lambda, U(\lambda))$ in Figure 1a. Under event E_δ and the separability condition, Lemma 7 and Theorem 10 give a lower bound on $\widehat{d}(\lambda)$ for $\lambda \geq \lambda^*$ and is denoted by the line $(\lambda, L(\lambda))$ in Figure 1a. These two bounds immediately give upper and lower bounds on the value thresholding estimator $\widehat{\lambda}^V_\tau$ for any $\tau \in [0, \infty)$. An illustration is provided in Figure 1a by the horizontal line through $(1, \tau)$. The points of intersection of this line with the feasible values of $(\lambda, \widehat{d}(\lambda))$ as in Figure 1a, given by r and s in the figure form lower and upper bounds respectively for $\widehat{\lambda}^V_\tau$.

□



(a) The feasible pairs of $(\lambda, \widehat{d}(\lambda))$ is shaded in light cyan.



(b) The feasible pairs of $(\lambda, \nabla\widehat{d}(\lambda))$ is shaded in light cyan.

Figure 1. Illustration of the upper and lower bounds on $\widehat{d}(\lambda)$ and $\nabla\widehat{d}(\lambda)$, under separability conditions (with margin α and tolerance β) and event E_δ.

Theorem 13. *Let $k(x, x) \leq 1$ for all $x \in \mathcal{X}$. Let the kernel k, and distributions G, H satisfy the separability condition with tolerance β and margin $\alpha > 0$. Let $\nu \in [\frac{\alpha}{4\lambda^*}, \frac{3\alpha}{4\lambda^*}]$ and $\sqrt{\min(m, n)} \geq \frac{36\sqrt{\log(1/\delta)}}{\frac{\alpha}{\lambda^*} - \nu}$. We then have with probability $1 - 4\delta$*

$$\lambda^* - \widehat{\lambda}^G_\nu \leq c \cdot \sqrt{\log(1/\delta)} \cdot (\min(m, n))^{-1/2},$$
$$\widehat{\lambda}^G_\nu - \lambda^* \leq \frac{4\beta\lambda^*}{\alpha} + c' \cdot \sqrt{\log(1/\delta)} \cdot (\min(m, n))^{-1/2},$$

for constants $c = (2\lambda^* - 1 + \sqrt{2\lambda^*}) \cdot \frac{12\lambda^*}{\alpha}$ and $c' = \frac{144(\lambda^*)^2(\alpha + 4\beta)}{\alpha^2}$.

Proof. (Sketch) The upper and lower bounds on $\widehat{d}(\lambda)$ given by Lemmas 7, 6 and Theorem 10 also immediately translate into upper and lower bounds for $\nabla\widehat{d}(\lambda)$ (assume differentiability of $\widehat{d}(.)$ for convenience) due to convexity of $\widehat{d}(.)$. As shown in Figure 1a, the gradient of $\widehat{d}(.)$ at some $\lambda_1 < \lambda^*$ is upper bounded by the slope of the line joining $(\lambda_1, 0)$ and $(\lambda^*, U(\lambda^*))$. Similarly, the gradient of $\widehat{d}(.)$ at some $\lambda_2 > \lambda^*$ is lower bounded by the slope of the line joining $(\lambda^*, U(\lambda^*))$ and $(\lambda_2, L(\lambda_2))$. Along with trivial bounds on $\nabla\widehat{d}(\lambda)$, these bounds give the set of feasible values for the ordered pair $(\lambda, \nabla\widehat{d}(\lambda))$, as illustrated in Figure 1b. This immediately gives bounds on $\widehat{\lambda}^G_\nu$ for any $\nu \in [0, \infty)$. An illustration is provided in Figure 1b by the horizontal line through $(1, \nu)$. The points of intersection of this line with the feasible values of $(\lambda, \nabla\widehat{d}(\lambda))$ as in Figure 1b, given by t and u in the figure form lower and upper bounds respectively for $\widehat{\lambda}^G_\nu$. □

Remark: Both the value and gradient thresholding estimates converge to λ^* with rates $O(m^{-\frac{1}{2}})$, if the kernel satisfies the separability condition with a tolerance $\beta = 0$. In the event of the kernel only satisfying the separability condition with tolerance $\beta > 0$, the estimates converge to within an additive factor of $\frac{\beta\lambda^*}{\alpha}$. As shown in Theorem 11, with a universal kernel the ratio $\frac{\beta}{\alpha}$ can be made arbitrarily low, and hence both the estimates actually converge to λ^*, but a specific rate is not possible, due to the dependence of the constants on α and β, without further assumptions on G and H.

6. The Gradient Thresholding Algorithm

As can be seen in Theorems 12 and 13, the value and gradient thresholding estimators both converge to λ^* at a rate of $O(m^{-\frac{1}{2}})$, in the scenario where we know the optimal threshold. In practice, one needs to set the threshold heuristically, and we observe that the estimate $\widehat{\lambda}^V_\tau$ is much more sensitive to the threshold τ, than the gradient thresholding estimate $\widehat{\lambda}^G_\nu$ is to the threshold ν. This agrees with our intuition of the asymptotic behavior of $\widehat{d}(\lambda)$ and $\nabla\widehat{d}(\lambda)$ – the curve of $\widehat{d}(\lambda)$ vs λ is close to a hinge, whereas the curve of $\nabla\widehat{d}(\lambda)$ vs λ is close to a step function. This can also be seen in Figure 2b. Hence, our estimator of choice is the gradient thresholding estimator and we give an algorithm for implementing it in this section.

Due to the convexity of $\widehat{d}(.)$, the slope $\nabla\widehat{d}(.)$ is an increasing function, and thus the gradient thresholding estimator $\widehat{\lambda}^G_\nu$ can be computed efficiently via binary search. The details of the computation are given in Algorithm 1.

Algorithm 1 maintains upper and lower bounds (λ_{left} and λ_{right}) on the gradient thresholding estimator,[1] estimates the slope at the current point λ_{curr} and adjusts the upper and

[1] We assume an initial upper bound of 10 for convenience, as we don't gain much by searching over higher values. $\widehat{\lambda}^G_\nu = 10$ corresponds to a mixture proportion estimate of $\widehat{\kappa} = 0.9$.

Algorithm 1 Kernel mean based gradient thresholder

1: **Input:** $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ drawn from mixture F and $\mathbf{x}_{n+1}, \ldots, \mathbf{x}_{n+m}$ drawn from component H
2: **Parameters:** $k : \mathcal{X} \times \mathcal{X} \to [0, \infty)$, $\nu \in [0, \infty)$
3: **Output:** $\widehat{\lambda}_\nu^G$
4: **Constants:** $\epsilon = 0.04, \lambda_{\text{UB}} = 10$
5: $\lambda_{\text{left}} = 1, \lambda_{\text{right}} = \lambda_{\text{UB}}$
6: $K_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j)$ for $1 \leq i, j \leq n + m$
7: **while** $\lambda_{\text{right}} - \lambda_{\text{left}} \geq \epsilon$
8: $\quad \lambda_{\text{curr}} = \frac{\lambda_{\text{right}} + \lambda_{\text{left}}}{2}$
9: $\quad \lambda_1 = \lambda_{\text{curr}} - \epsilon/4$
10: $\quad \mathbf{u}_1 = \frac{\lambda_1}{n}([\mathbf{1}_n^\top, \mathbf{0}_m^\top]) + \frac{1-\lambda_1}{m}([\mathbf{0}_n^\top, \mathbf{1}_m^\top])$
11: $\quad d_1 = \widehat{d}(\lambda_1)^2 = \min_{\mathbf{v} \in \Delta_{n+m}} (\mathbf{u}_1 - \mathbf{v})^\top K (\mathbf{u}_1 - \mathbf{v})$
12: $\quad \lambda_2 = \lambda_{\text{curr}} + \epsilon/4$
13: $\quad \mathbf{u}_2 = \frac{\lambda_2}{n}([\mathbf{1}_n^\top, \mathbf{0}_m^\top]) + \frac{1-\lambda_2}{m}([\mathbf{0}_n^\top, \mathbf{1}_m^\top])$
14: $\quad d_2 = \widehat{d}(\lambda_2)^2 = \min_{\mathbf{v} \in \Delta_{n+m}} (\mathbf{u}_2 - \mathbf{v})^\top K (\mathbf{u}_2 - \mathbf{v})$
15: $\quad s = \frac{\sqrt{d_2} - \sqrt{d_1}}{\lambda_2 - \lambda_1}$
16: $\quad$ **if** $s > \nu$:
17: $\quad\quad \lambda_{\text{right}} = \lambda_{\text{curr}}$
18: $\quad$ **else**:
19: $\quad\quad \lambda_{\text{left}} = \lambda_{\text{curr}}$
20: **return** λ_{curr}

lower bounds based on the computed slope. The slope at the current point λ_{curr} is estimated numerically by computing the value of $\widehat{d}(.)$ at $\lambda_{\text{curr}} \pm \frac{\epsilon}{4}$ (lines 9 to 15). We compute the value of $\widehat{d}(\lambda)$ for some given λ using the general purpose convex programming solver CVXOPT. [2]

We employ the following simple strategy for model selection (choosing the kernel k and threshold ν). Given a set of kernels, we choose the kernel for which the empirical RKHS distance between the distributions F and H, given by $\|\phi(\widehat{F}) - \phi(\widehat{H})\|_{\mathcal{H}}$ is maximized. This corresponds to choosing a kernel for which the "roof" of the step-like function $\nabla\widehat{d}(.)$ is highest. We follow two different strategies for setting the gradient threshold ν. One strategy is motivated by Lemma 6, where we can see that the slope of $\widehat{d}(\lambda)$ for $\lambda \in [1, \lambda^*]$ is $O(1/\sqrt{\min(m,n)})$ and based on this we set $\nu = 1/\sqrt{\min(m,n)}$. The other strategy is based on empirical observation, and is set as a convex combination of the initial slope of $\widehat{d}$ at $\lambda = 1$ and the final slope at $\lambda = \infty$ which is equal to the RKHS distance between the distributions F and H, given by $\|\phi(\widehat{F}) - \phi(\widehat{H})\|_{\mathcal{H}}$. We call the resulting two algorithms as "KM1" and "KM2" respectively in our experiments.[3]

[2] The accuracy parameter ϵ must be set large enough so that the optimization error in computing $\widehat{d}(\lambda_{\text{curr}} \pm \frac{\epsilon}{4})$ is small when compared to $\widehat{d}(\lambda_{\text{curr}} + \frac{\epsilon}{4}) - \widehat{d}(\lambda_{\text{curr}} - \frac{\epsilon}{4})$.

[3] In KM2, $\nu = 0.8 * \text{init\_slope} + 0.2 * \text{final\_slope}$

7. Other Methods for Mixture Proportion Estimation

Blanchard et al. (2010) propose an estimator based on the following equality, which holds under an irreducibility condition (which is a strictly weaker requirement than the anchor set condition), $\kappa^* = \inf_{\mathcal{S} \in \Theta, H(\mathcal{S}) > 0} \frac{F(\mathcal{S})}{H(\mathcal{S})}$, where Θ is the set of measurable sets in $\mathcal{X}$. The estimator proposed replaces the exact terms $F(\mathcal{S})$ and $H(\mathcal{S})$ in the above ratio with the empirical quantities $\widehat{F}(\mathcal{S})$ and $\widehat{H}(\mathcal{S})$ and includes VC-inequality based correction terms in the numerator and denominator and restricts Θ to a sequence of VC classes. Blanchard et al. (2010) show that the proposed estimator converges to the true proportion under the irreducibility condition and also show that the convergence can be arbitrarily slow. Note that the requirement of taking infimum over VC classes makes a direct implementation of this estimator computationally infeasible.

Scott (2015) show that the estimator of Blanchard et al. (2010) converges to the true proportion at the rate of $1/\sqrt{\min(m,n)}$ under the anchor set condition, and also make the observation that the infimum over the sequence of VC classes can be replaced by an infimum over just the collection of base sets (e.g. the set of all open balls). Computationally, this observation reduces the complexity of a direct implementation of the estimator to $O(N^d)$ where $N = m + n$ is the number of data points, and d is the data dimension. But the estimator still remains intractable for even datasets with moderately large number of features.

Sanderson & Scott (2014); Scott (2015) propose algorithms based on the estimator of Blanchard et al. (2010), which treats samples from F and samples from H as positive and negative classes, builds a conditional probability estimator and computes the estimate of κ^* from the constructed ROC (receiver operating characteristic) curve. These algorithms return the correct answer when the conditional probability function learned is exact, but the effect of error in this step is not clearly understood. This method is referred to as "ROC" in our experimental section.

Elkan & Noto (2008) propose another method for estimating κ^* by constructing a conditional probability estimator which treats samples from F and samples from H as positive and negative classes. Even in the limit of infinite data, it is known that this estimator gives the right answer only if the supports of G and H are completely distinct. This method is referred to as "EN" in our experiments.

du Plessis & Sugiyama (2014) propose a method for estimating κ^* based on Pearson divergence minimization. It can be seen as similar in spirit to the method of Elkan & Noto (2008), and thus has the same shortcoming of being exact only when the supports of G and H are disjoint, even in the limit of infinite data. The main difference between

Table 1. Dataset statistics

Dataset	# of samples	Pos. frac.	Dim.
waveform	3343	0.492	21
mushroom	8124	0.517	117
pageblocks	5473	0.897	10
shuttle	58000	0.785	9
spambase	4601	0.394	57
digits	13966	0.511	784

the two is that this method does not require the estimation of a conditional probability model as an intermediate object, and computes the mixture proportion directly.

Recently, Jain et al. (2016) have proposed another method for the estimation of mixture proportion which is based on maximizing the "likelihood" of the mixture proportion. The algorithm suggested by them computes a likelihood associated with each possible value of κ^*, and returns the smallest value for which the likelihood drops significantly. In a sense, it is similar to our gradient thresholding algorithm, which also computes a distance associated to each possible value of λ^*, and returns the smallest value for which the distance increases faster than a threshold. Their algorithm also requires a conditional probability model distinguishing F and H to be learned. It also has no guarantees of convergence to the true estimate κ^*. This method is referred to as "alphamax" in our experiments.

Menon et al. (2015); Liu & Tao (2015) and Scott et al. (2013b) propose to estimate the mixture proportion κ^*, based on the observation that, if the distributions F and H satisfy the anchor set condition, then κ^* can be directly related to the maximum value of the conditional probability given by $\max_x \eta(x)$, where η is the conditional probability function in the binary classification problem treating samples from F as positive and samples from H negative. Thus one can get an estimate of κ^* from an estimate of the conditional probability $\widehat{\eta}$ through $\max_x \widehat{\eta}(x)$. This method clearly requires estimating a conditional probability model, and is also less robust to errors in estimating the conditional probability due to the form of the estimator.

8. Experiments

We ran our algorithm with 6 standard binary classification datasets[4] taken from the UCI machine learning repository, the details of which are given below in Table 1.[5]

From each binary dataset containing positive and negative labelled data points, we derived 6 different pairs of mixture and component distributions (F and H respectively) as follows. We chose a fraction of the positive data points to be part of the component distribution, the positive data points not chosen and the negative data points constitute the mixture distribution. The fraction of positive data points chosen to belong to the component distribution was one of $\{0.25, 0.5, 0.75\}$ giving 3 different pairs of distributions. The positive and negative labels were flipped and the above procedure was repeated to get 3 more pairs of distributions. From each such distribution we drew a total of either 400,800,1600 or 3200 samples and ran the two variants of our kernel mean based gradient thresholding algorithm given by "KM1" and "KM2". Our candidate kernels were five Gaussian RBF kernels, with the kernel width taking values uniformly in the log space between a tenth of the median pairwise distance and ten times the median distance, and among these kernels the kernel for which $\|\phi(\widehat{F}) - \phi(\widehat{H})\|$ is highest is chosen. We also ran the "alphamax", "EN" and "ROC" algorithms for comparison.[6] The above was repeated 5 times with different random seeds, and the average error $|\widehat{\kappa} - \kappa^*|$ was computed. The results are plotted in Figure 3 and the actual error values used in the plots is given in the supplementary material Section H. Note that points in all plots are an average of 30 error terms arising from the 6 distributions for each dataset, and 5 different sets of samples for each distribution arising due to different random seeds.

It can be seen from the plots in Figure 3, that our algorithms (KM1 and KM2) perform comparably to or better than other algorithms for all datasets except mushroom.

9. Conclusion

Mixture proportion estimation is an interesting and important problem that arises naturally in many 'weakly supervised learning' settings. In this paper, we give an efficient kernel mean embedding based method for this problem, and show convergence of the algorithm to the true mixture proportion under certain conditions. We also demonstrate the effectiveness of our algorithm in practice by running it on several benchmark datasets.

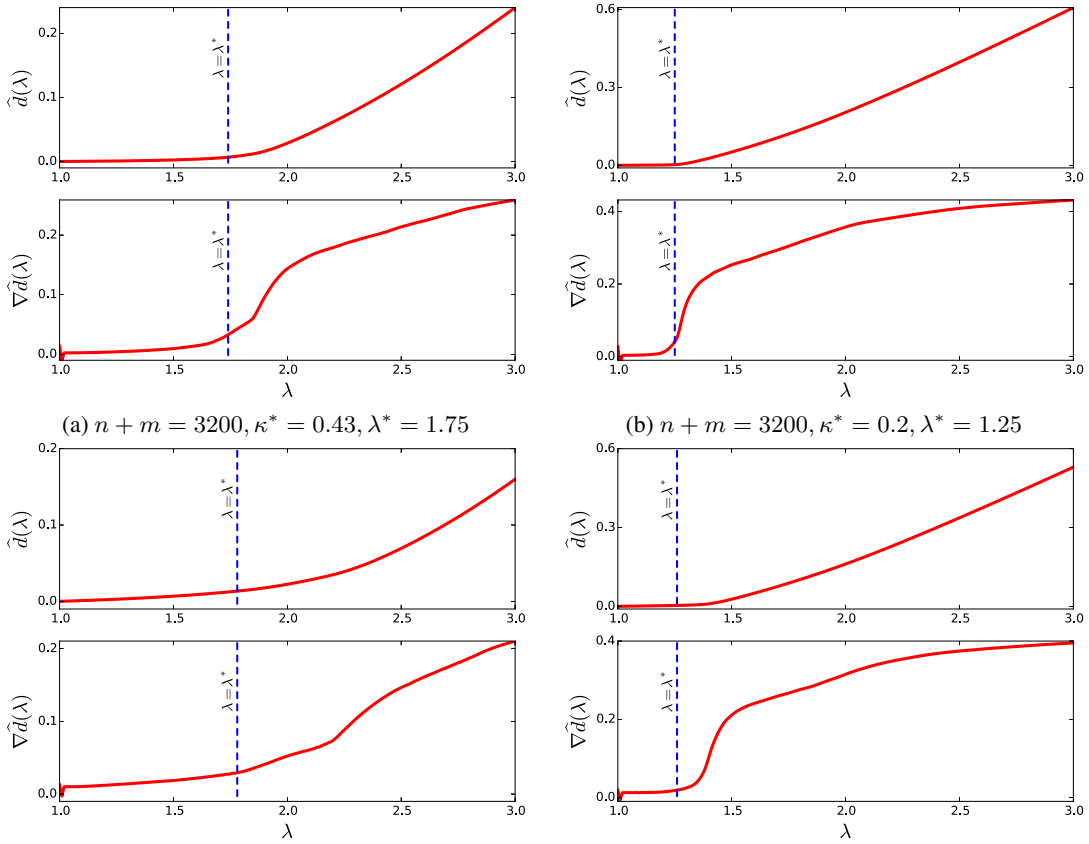
Acknowledgements

This work was supported in part by NSF Grants No. 1422157, 1217880, and 1047871.

[4] shuttle, pageblocks, digits are originally multiclass datasets, they are used as binary datasets by either grouping or ignoring classes.

[5] In our experiments, we project the data points from the digits and mushroom datasets onto a 50-dimensional space given by PCA.

[6] The code for our algorithms KM1 and KM2 are at http://web.eecs.umich.edu/~cscott/code.html#kmpe. The code for ROC was taken from http://web.eecs.umich.edu/~cscott/code/mpe.zip. The codes for the alphamax and EN algorithms were the same as in Jain et al. (2016), and acquired through personal communication.



(a) $n+m=3200, \kappa^*=0.43, \lambda^*=1.75$

(b) $n+m=3200, \kappa^*=0.2, \lambda^*=1.25$

(c) $n+m=800, \kappa^*=0.43, \lambda^*=1.75$

(d) $n+m=800, \kappa^*=0.2, \lambda^*=1.25$

Figure 2. $\widehat{d}(.)$ and $\nabla\widehat{d}(.)$ are plotted for two different sample sizes and true positive proportions.

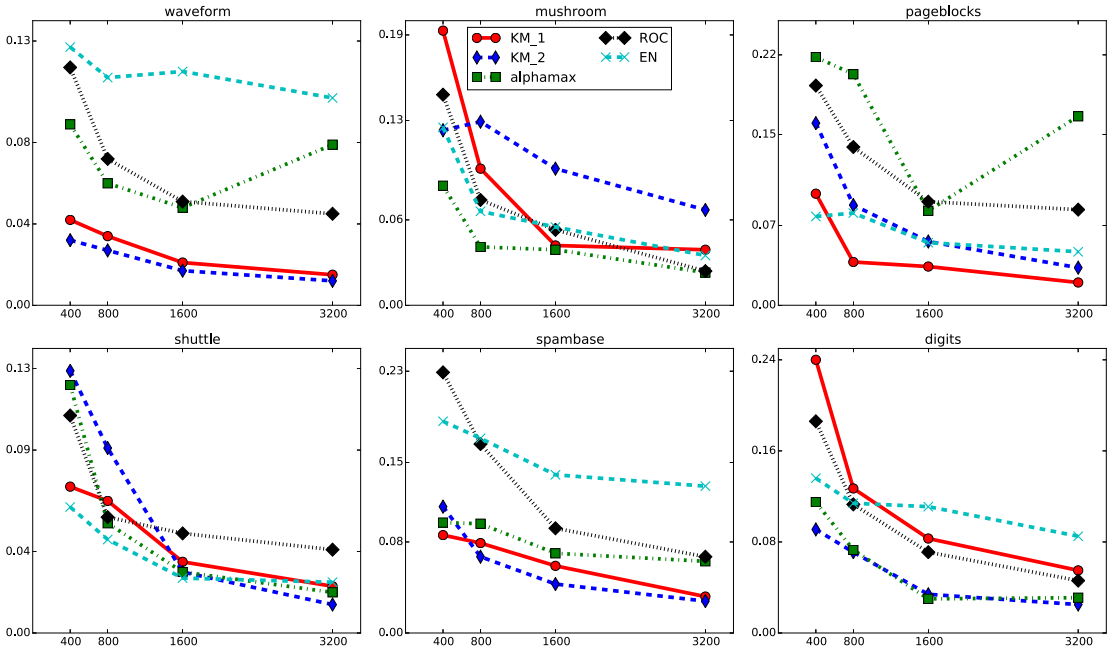


Figure 3. The average error made by the KM, alphamax, ROC and EN algorithms in predicting the mixture proportion κ^* for various datasets as a function of the total number of samples from the mixture and component.

References

Aronszajn, N. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950.

Arora, S., Ge, R., and Moitra, A. Learning topic models – going beyond SVD. In *Proceedings of IEEE Foundations of Computer Science (FOCS)*, pp. 1–10, 2012.

Berlinet, A. and Thomas, C. *Reproducing kernel Hilbert spaces in Probability and Statistics*. Kluwer Academic Publishers, 2004.

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

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

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

du Plessis, M. C. and Sugiyama, M. Class prior estimation from positive and unlabeled data. *IEICE Transactions on Information and Systems*, 97:1358–1362, 2014.

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

Gretton, A., Borgwardt, K. M., Rasch, M. J., Scholkopf, B., and Smola, A. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

Jain, S., White, M., Trosset, M. W., and Radivojac, P. Nonparametric semi-supervised learning of class proportions. *arXiv:1601.01944*, 2016.

Lawrence, N. and Scholkopf, B. Estimating a kernel Fisher discriminant in the presence of label noise. In *Proc. of the Int. Conf. in Machine Learning (ICML)*, 2001.

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

Liu, T. and Tao, D. Classification with noisy labels by importance reweighting. *IEEE Transactions on pattern analysis and machine intelligence*, 38(3):447–461, 2016.

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

Menon, A. K., van Rooyen, B., Ong, C. S., and Williamson, R. C. Learning from corrupted binary labels via class-probability estimation. In *In Proc. of the Int. Conf. in Machine Learning (ICML)*, pp. 125–134, 2015.

Michelli, C., Xu, Y., and Zhang, H. Universal kernels. *Journal of Machine Learning Research*, 7:2651–2667, 2006.

Natarajan, N., Dhillon, I. S., Ravikumar, P., and Tewari, A. Learning with noisy labels. In *Advances in Neural Information Processing Systems (NIPS) 26*, pp. 1196–1204, 2013.

Raykar, V. C., Yu, S., Zhao, L. H., Valadez, G. H., Florin, C., Bogoni, L., and Moy, L. Learning from crowds. *The Journal of Machine Learning Research*, 11:1297–1322, 2010.

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

Scott, C. A rate of convergence for mixture proportion estimation, with application to learning from noisy labels. In *Proc. of the Int. Conf. on Artificial Intelligence and Statistics (AISTATS)*, 2015.

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

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

Smola, A., Gretton, A., Song, L., and Scholkopf, B. A Hilbert space embedding for distributions. In *Algorithmic Learning Theory (ALT)*, 2007.

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

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