An Efficient and Provable Approach for Mixture Proportion Estimation Using Linear Independence Assumption

Xiyu Yu[1] Tongliang Liu[1] Mingming Gong[2,3] Kayhan Batmanghelich[2] Dacheng Tao[1]
[1]UBTECH Sydney AI Centre, SIT, FEIT, The University of Sydney, Australia
[2]Department of Biomedical Informatics, University of Pittsburgh
[3]Department of Philosophy, Carnegie Mellon University
{xiyu0300@uni., tongliang.liu@, dacheng.tao@}sydney.edu.au {mig73, kayhan}@pitt.edu

Abstract

In this paper, we study the mixture proportion estimation (MPE) problem in a new setting: given samples from the mixture and the component distributions, we identify the proportions of the components in the mixture distribution. To address this problem, we make use of a linear independence assumption, i.e., the component distributions are independent from each other, which is much weaker than assumptions exploited in the previous MPE methods. Based on this assumption, we propose a method (1) that uniquely identifies the mixture proportions, (2) whose output provably converges to the optimal solution, and (3) that is computationally efficient. We show the superiority of the proposed method over the state-of-the-art methods in two applications including learning with label noise and semi-supervised learning on both synthetic and real-world datasets.

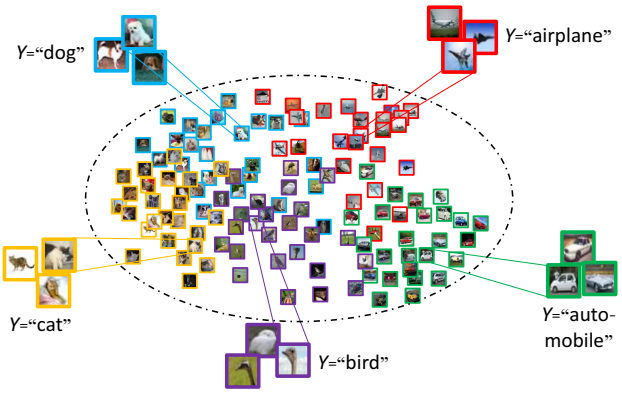


Figure 1. An illustration of the setting in this study. We are provided with unlabeled or noisy data sampled from the mixture distribution (images inside the ellipse with black dashed line) and also a small set of correctly labeled data sampled from each component distribution (images outside the ellipse). Here, images with different color boundaries come from different classes. We aim to estimate the weights of components based on these information.

1. Introduction

The estimation of the proportions of component distributions in a mixture, namely, mixture proportion estimation (MPE), has been an important prerequisite for many practical problems. For example, in the label noise problem where training data are randomly mislabelled with some small flip probabilities [9], each class-conditional distribution of noisy data is a mixture of the true class-conditional distributions. The mixture proportions, which are closely related to the flip rates, are essential for designing noise-robust loss functions [31, 22, 24], and can be estimated by MPE methods. MPE also arises in the scenario of semi-supervised learning. The proportions of positive and negative examples in the unlabelled sample are often required to design cost-sensitive loss functions [23, 29].

If we are only given data from the mixture distribution, we can possibly estimate the mixture proportions and component distributions by unsupervised learning [27, 3]. But these unsupervised learning methods typically rely on strong restrictions on the distribution class, for example, mixture of Gaussians (MoG), and the solution is generally not unique. Thus, existing MPE methods focus on the weakly-supervised setting, e.g., in a mixture distribution containing two component distributions, samples from the mixture and one component are available. This setting has found applications in PU learning (learning from positive and unlabeled examples) [8] and multi-instance learning [26]. Still, a unique solution cannot be guaranteed in this setting without making further assumptions.

Two popular assumptions, including irreducibility and anchor set conditions, have been introduced to make the problem tractable. Here, we say a component is the target component if we want to estimate its proportion in the mixture. The irreducibility condition assumes that the tar-

get component cannot be represented by a convex combination of the remaining components and a new distribution [2, 32]. Under this condition, the proportion of the target component can be provably identifiable. However, uniform convergence of the estimator in [2] is not guaranteed, and can be arbitrarily slow in practice. The anchor set condition [31, 26] is another stronger assumption. It assumes that each component has a compact support set not shared with others, and this compact support set is called the anchor set. This assumption leads to estimation methods that can provably converge to the true proportion with a guaranteed rate.

However, all of the aforementioned methods suffer from one or more of the following weaknesses. First, although the irreducibility and anchor set assumptions are necessary for proving identifiability or convergence rates, they can easily be violated in practical problems. Second, the estimation error is usually large in practice. For example, methods like [31, 8, 19] rely on accurate estimation of the conditional probability, which is often unachievable because it is difficult to choose an appropriate model for the conditional distributions. Third, the estimation algorithms are inefficient. For example, methods proposed in [26] need to solve several quadratic programming problems, which is relatively time-consuming. Finally, the extension of these methods to more than two mixture components is not straightforward.

In some problems such as semi-supervised learning, we have access to a few examples from every component distribution and sufficient examples from the mixture distribution, as shown in Figure 1. To target this kind of problems, we study the MPE problem in a new setting: we estimate the proportions of the component distributions in a mixture given the samples from the mixture and all components. We find that a much weaker assumption, i.e., the independence of component distribution, is sufficient to guarantee the identifiability of mixture proportions. Under this assumption, we propose an estimation method by embedding the distributions into a reproducing kernel Hilbert space (RKHS). Our method only requires solving a simple quadratic programming problem, which is much more efficient than previous methods. Furthermore, under such a weaker assumption, the proposed method also gives a consistent estimator which provably converges to the true proportion with a guaranteed convergence rate.

We demonstrate the effectiveness of our MPE method in two applications including flip rate estimation in learning with label noise and class prior estimation in semi-supervised learning. We validate our approach with comprehensive experiments. Unlike many previous flip rate estimation methods, our method can be easily applied to the multi-class setting, and enjoy the efficiency and high performances on both synthetic and real-world datasets.

2. Related Work

2.1. Mixture proportion estimation

Most of the recent MPE methods are based on the anchor set condition. For instance, [30, 31] proposed the ROC (receiver operating characteristic) curve-based methods, which can provably converge to the true proportion with a guaranteed rate. [19, 24] proposed to estimate the proportions based on the examples in the anchor set. [26] proposed a kernel mean based gradient thresholding algorithm to identify the proportion of a component distribution from a mixture. The distributions are first embedded into a RKHS, and then the optimal proportion is greedily searched in an interval by solving a series of quadratic programming problems. Under the anchor set condition, this method can provably converge to the true proportion.

Since [26] is the closest work to ours, we summarize the main differences from two aspects. First, we explore the use of the independence assumption rather than the anchor set one in mixture proportion estimation. The different assumptions of component distributions result in distinct analysis of the convergence rate. [26] provided a convergence rate with the order of $O(1/\sqrt{\min(n, n_0)})$, where n and n_0 are the sample sizes of the data from the mixture and the target component, respectively. However, the estimate converges to within an additive factor of the true proportion, which heavily depends on the choice of the kernel function and the property of the anchor set. This factor may lead to a slow convergence in practice. However, in this paper, even though the proved upper bound has a relatively slower convergence rate, the estimates are guaranteed to converge to the optimal solutions.

Second, [26] considered only a mixture of two components given the samples from the mixture and the target component. In this paper, we study a more general case, that is, a mixture of multiple components, where the samples from the mixture and all components are given. Even though our model requires a sample for each component, having a small number of examples for each component is empirically sufficient, which are easily obtainable in many problems. This setting has applications in many problems, such as learning with label noise [37, 18] and semi-supervised learning [29].

2.2. Class ratio estimation

Class ratio estimation [7, 14, 6] studies a similar problem that estimates the class ratios of unlabeled data given a small set of labeled training data. Iyer et al. [14] exploited the maximum mean discrepancy (MMD) framework to solve the class ratio estimation problem, which shares a similar formulation with the proposed method. But to the best of our knowledge, we are the first to comprehensively study the relationships between several assumptions

of MPE and propose to solve MPE by using the weakest linear independence assumption. Furthermore, we prove a novel convergence bound which is data-independent; that is to say, the proposed method can uniformly converge to the optimal solution. On the other hand, the error bound provided by Iyer et al. [14] is proportional to the reciprocal of the minimum eigenvalue of a data-dependent matrix. If the training data are similar, the minimum eigenvalue can be as small as $\frac{1}{n}$, which leads to the fact that their convergence rate can be arbitrary slow. Furthermore, we have explicitly studied the relationship between the independence of distributions and the independence of kernel mean embeddings. Finally, we have studied a wide range of applications of our MPE method and have contributed to learning with multi-class label noise and semi-supervised learning.

3. MPE with Linear Independence Assumption

Suppose that $P_i, i \in \{1, \cdots, c\}$, are $c \geq 2$ different component distributions over a compact metric space $\mathcal{X}$, and that P is their mixture, which satisfies

$$P = \sum_{i=1}^{c} \lambda_i P_i, \tag{1}$$

where $\lambda_i \geq 0, \forall i \in \{1, \cdots, c\}$ and $\sum_{i=1}^{c} \lambda_i = 1$. We consider the setting that samples from all components $\{x_1^i, x_2^i, \cdots, x_{n_i}^i\}$, $i = 1, \cdots, c$, and the sample from the mixture distribution $\{x_1, x_2, \cdots, x_n\}$ are given. Mixture proportion estimation (MPE) aims to estimate $\{\lambda_i\}_{1 \leq i \leq c}$ from the data. Here $n, n_1, \cdots, n_c$ are the sample sizes of the mixture and the c components, respectively.

3.1. Linear independence assumption

Without any assumption on the component distributions, it is not possible to identify the proportions. Previous methods assume the irreducibility and anchor set conditions, which are formally defined in the following two definitions. Without loss of generality, we consider the case of two component distributions, that is, the mixture $P = \lambda_1 P_1 + \lambda_2 P_2$, where $\lambda_1 \geq 0, \lambda_2 \geq 0$ and $\lambda_1 + \lambda_2 = 1$.

Definition 1 (Mutual Irreducibility)**.** *Distribution P_1 is irreducible to P_2 if there exist **no** such $\gamma \in (0, 1]$ and any new distribution Q that*

$$P_1 = \gamma P_2 + (1 - \gamma) Q.$$

If P_2 is also irreducible to P_1, then they are mutual irreducible.

Definition 2 (Anchor Set)**.** *Denote $supp(\cdot)$ as the support set of a distribution, distributions P_1 and P_2 satisfy the anchor set condition if there exist a compact set S, which is non-empty, such that $S \subseteq (supp(P_1) \cup supp(P_2) - supp(P_1) \cap supp(P_2))$.*

In this paper, with the minor requirement of some examples from each component, our method only needs a much weaker independence assumption to ensure efficient and effective solutions.

Assumption 1. *The component distributions P_i are independent from each other, that is,*

$$\sum_{i=1}^{c} v_i P_i = 0, v_i \in \mathbb{R} \text{ if and only if } v_i = 0, \forall i = \{1, \cdots, c\}.$$

Considering probability densities as functions in an infinite dimension space, our definition is closely related to the independence assumption in a linear algebraic sense. The linear independence assumption can be easily satisfied in practice because the number of component distributions is finite. Moreover, the following proposition states that the independence assumption is weaker than the irreducibility and anchor set conditions.

Proposition 1. *(i) The irreducibility condition implies the independence assumption while the independence assumption does not imply the irreducibility condition.*

(ii) The anchor set condition implies the independence assumption while the independence assumption does not imply the anchor set condition.

The proof of Proposition 1 can be found in the supplementary material. In the following theorem, we will show that the proportions are identifiable under the independence assumption.

Theorem 1. *If Assumption 1 holds, the mixture proportions λ_i are identifiable given the mixture distribution and all the component distributions.*

The proof of Theorem 1 is straightforward. Suppose there exist another mixture proportions $\{\lambda_i'\}_{1 \leq i \leq c}$ that admit Eq. (1), which means $P = \sum_{i=1}^{c} \lambda_i' P_i$, then we have $\sum_{i=1}^{c} (\lambda_i - \lambda_i') P_i = 0$. If Assumption 1 holds true, then $\lambda_i - \lambda_i' = 0, \ \forall i = \{1, \cdots, c\}$. This means the two proportions are identical. Thus, the solution to mixture proportions in Eq. (1) is unique.

3.2. MPE model

In order to estimate $\{\lambda_i\}_{1 \leq i \leq c}$ without estimating the conditional probability as in [31, 19], we propose a non-parametric method which is based on the embedding of the distributions into a reproducing kernel Hilbert space $\mathcal{H}$ [33, 21]. Let P_X be the distribution of the variable $X \in \mathcal{X}$, $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be the kernel function associated with $\mathcal{H}$, and $k(X, \cdot) = \psi(X) \in \mathcal{H}$ be the feature map. The kernel mean embedding is defined as

$$\mu_{P_X} = \mathbb{E}_{X \sim P_X}[k(X, \cdot)],$$

where μ_{P_X} is known as the kernel mean embedding of P_X. The mean map is injective given the kernel is characteristic [34]. We can thus estimate $\{\lambda_i\}_{1\leq i\leq c}$ by exploiting these kernel means. Using linearity of the expectation, one can re-write the mixture form the corresponding kernel mean

$$\mu_P = \sum_{i=1}^{c} \lambda_i \mu_{P_i}. \tag{2}$$

As demonstrated in Theorem 1, if these kernel means in Eq. (2) are independent, then the uniqueness of $\{\lambda_i\}_{1\leq i\leq c}$ is also ensured. In the following theorem, we will show that the kernel means are independent given that the component distributions are independent.

Theorem 2. *Suppose that the kernel is characteristic, and that distributions $P_i, i = 1, \cdots, c$ satisfy Assumption 1, then we have*

$$\sum_{i=1}^{c} v_i \mu_{P_i} = 0, v_i \in \mathbb{R} \text{ if and only if } v_i = 0, \forall i = \{1, \cdots, c\}.$$

Proof. A necessary condition for a kernel to be characteristic is $\int k(x,\cdot)dP_X = 0 \Rightarrow P_X = 0$ [10]. Then $\sum_{i=1}^{c} v_i \mu_{P_i} = \sum_{i=1}^{c} v_i \int k(x,\cdot)dP_i = \int k(x,\cdot)d(\sum_{i=1}^{c} v_i P_i) = 0 \Rightarrow \sum_{i=1}^{c} v_i P_i = 0$. Since $P_i, i \in \{1, \cdots, c\}$ are independent, we have $\sum_{i=1}^{c} v_i P_i = 0 \Rightarrow v_i = 0, \forall i \in \{1, \cdots, c\}$, which means $\mu_{P_i}, i \in \{1, \cdots, c\}$ are also independent. □

According to Theorem 2, the independence assumption can finally ensure the **uniqueness** of the proportions. Under such a theoretical guarantee, we propose a method to estimate $\{\lambda_i\}_{1\leq i\leq c}$ from Eq. (2). We try to minimize the squared maximum mean discrepancy (MMD):

$$D(\boldsymbol{\lambda}) = \|\mu_P - \sum_{i=1}^{c} \lambda_i \mu_{P_i}\|^2,$$

which indicates $\mu_P = \sum_{i=1}^{c} \lambda_i \mu_{P_i}$ if $D(\boldsymbol{\lambda})$ is zero. With the guarantee of uniqueness, the proportions can thus be identifiable.

However, only the samples of these distributions are observable in practice. Therefore, we approximate the mean values by their empirical ones: $\hat{\mu}_P = \frac{1}{n}\sum_{j=1}^{n} \psi(x_j)$, and $\hat{\mu}_{P_i} = \frac{1}{n_i}\sum_{j=1}^{n_i} \psi(x_j^i), i = 1, \cdots, c$. Then the problem becomes

$$\min_{\lambda_1,\cdots,\lambda_c} \hat{D}(\boldsymbol{\lambda}) = \|\frac{1}{n}\sum_{j=1}^{n} \psi(x_j) - \sum_{i=1}^{c} \frac{\lambda_i}{n_i} \sum_{j=1}^{n_i} \psi(x_j^i)\|^2,$$
$$s.t. \quad \lambda_i \geq 0, \forall i \in \{1, \cdots, c\} \text{ and } \sum_{i=1}^{c} \lambda_i = 1. \tag{3}$$

Let $m = \sum_{i=1}^{c} n_i$ be the total number of examples from all component distributions and d denotes the dimensionality of the data. Then the data matrix composed of data points from the mixture distribution can be denoted as $\mathbf{x}^M \in \mathbb{R}^{n\times d}$, and the data matrix containing samples from all the component distributions is denoted as $\mathbf{x}^C \in \mathbb{R}^{m\times d}$. We can write the kernel mean embeddings in matrix forms. $\frac{1}{n}\sum_{j=1}^{n} \psi(x_j) = \frac{1}{n}\psi(\mathbf{x}^M)^\top \mathbf{1}$, where $\mathbf{1}$ is an all-ones vector. $\sum_{i=1}^{c} \frac{\lambda_i}{n_i}\sum_{j=1}^{n_i} \psi(x_j^i) = \psi(\mathbf{x}^C)R\boldsymbol{\lambda}$, where $\boldsymbol{\lambda} = [\lambda_1, \cdots, \lambda_c]^\top$ and $R \in \mathbb{R}^{m\times c}$. $R_{ij} = \frac{1}{n_j}$ if the i-th example is sampled from distribution P_j; otherwise, $R_{ij} = 0$. Then the objective function can be rewritten as

$$\begin{aligned}\hat{D}(\boldsymbol{\lambda}) &= \|\frac{1}{n}\psi(\mathbf{x}^M)^\top \mathbf{1}_n - \psi(\mathbf{x}^C)R\boldsymbol{\lambda}\|^2 \\ &= \boldsymbol{\lambda}^\top R^\top \mathbf{K}^C R\boldsymbol{\lambda} - \frac{2}{n}\mathbf{1}^\top \mathbf{K}^{M,C} R\boldsymbol{\lambda} + \frac{1}{n^2}\mathbf{1}^\top \mathbf{K}^M \mathbf{1},\end{aligned}$$

where $\mathbf{K}^M$ and $\mathbf{K}^C$ are the kernel matrix of $\mathbf{x}^M$ and $\mathbf{x}^C$, respectively; $\mathbf{K}^{M,C}$ is the cross-kernel matrix. In this paper, the Gaussian kernel, i.e. $k(x_i, x_j) = \exp(-\frac{\|x_i - x_j\|^2}{2\gamma^2})$ is applied, where γ is the kernel bandwidth. Then our final model becomes

$$\min_{\lambda_1,\cdots,\lambda_c} \boldsymbol{\lambda}^\top R^\top \mathbf{K}^C R\boldsymbol{\lambda} - \frac{2}{n}\mathbf{1}^\top \mathbf{K}^{M,C} R\boldsymbol{\lambda} + \frac{1}{n^2}\mathbf{1}^\top \mathbf{K}^M \mathbf{1},$$
$$s.t. \quad \lambda_i \geq 0, \forall i \in \{1, \cdots, c\} \text{ and } \sum_{i=1}^{c} \lambda_i = 1.$$

This is a convex quadratic programming problem, which can be easily solved using standard procedures. We can see that, compared to the traditional MPE methods, our method avoids the estimation of the conditional probability and saves computations as well.

3.3. Theoretical analysis

Given samples $\{x_1^i, x_2^i, \cdots, x_{n_i}^i\}$, $i = 1, \cdots, c$, drawn from the components, and the sample $\{x_1, x_2, \cdots, x_n\}$ from the mixture distribution, we can obtain an estimate $\hat{\boldsymbol{\lambda}}$ by solving model (3). An important issue we are concerned about is how quickly $\hat{\boldsymbol{\lambda}}$ can converge to the optimal one $\boldsymbol{\lambda}^*$, where $\boldsymbol{\lambda}^* = \arg\min_{\lambda_1,\cdots,\lambda_c} D(\boldsymbol{\lambda})$. In this section, under the independence assumption, we deliver a convergence analysis of the proposed algorithm.

We abuse the samples $\{x_1^i, x_2^i, \cdots, x_{n_i}^i\}$, $i = 1, \cdots, c$ and $\{x_1, x_2, \cdots, x_n\}$ as being i.i.d. variables, then $D(\boldsymbol{\lambda})$ can be rewritten as

$$D(\boldsymbol{\lambda}) = \left\| \mathbb{E}\left[\frac{1}{n}\sum_{j=1}^{n} \psi(x_j) - \sum_{i=1}^{c} \frac{\lambda_i}{n_i} \sum_{j=1}^{n_i} \psi(x_j^i) \right] \right\|^2.$$

Due to the fact that when $\hat{\boldsymbol{\lambda}}$ approaches to $\boldsymbol{\lambda}^*$, $D(\hat{\boldsymbol{\lambda}})$ also converges to $D(\boldsymbol{\lambda}^*)$. We can thus analyze the convergence rate of $\boldsymbol{\lambda}$ by upper bounding the error $D(\hat{\boldsymbol{\lambda}}) - D(\boldsymbol{\lambda}^*)$. Here comes to our main result,

Theorem 3. *Suppose the kernel is characteristic and upper bounded by* $\|\psi(x)\|_2 \leq r$ *for all* $x \in \mathcal{X}$. *Then for any* $\delta > 0$, *with the probability* $1-\delta$, *we have*

$$D(\hat{\boldsymbol{\lambda}}) - D(\boldsymbol{\lambda}^*) \leq 8\sqrt{2}r^2 \sqrt{(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{n_0}}) + \sqrt{\frac{1}{2}(\frac{1}{n} + \sum_{i=1}^{c} \frac{1}{n_i}) \log \frac{1}{\delta}}},$$

where $n_0 = \min(n_1, \cdots, n_c)$.

Please see the detailed proof of Theorem 3 in the supplementary material.

Remark 1. According to Theorem 3, our proposed algorithm converges to the optimal $\boldsymbol{\lambda}^*$ with rate of $O(1/\min(n^{\frac{1}{4}}, n_0^{\frac{1}{4}}))$. That is to say, to obtain an acceptable estimate, we only require a small number of examples drawn from the mixture and components.

Remark 2. [31] showed the ROC curve-based based estimator converges to the true proportion at the rate of $O(1/\min(\sqrt{n}, \sqrt{n_0}))$. However, as claimed in [26], this estimator cannot be applied to datasets with moderately large number of features. Thus, [26] proposed an estimator based on the kernel mean embedding, and proved a convergence rate with the order of $O(1/\min(\sqrt{n}, \sqrt{n_0}))$. However, it converges to the true proportion with an additive term, whose convergence rate is strongly dependent on the choice of the kernel function and the probability of the anchor set $P_i(S)$, and can be slow.

In this paper, under the independence assumption, even though with a slightly slower rate, the proposed method can be easily applied to datasets with large sample size and high dimensional features, and can be optimized efficiently by solving a simple quadratic programming problem.

4. Applications

Mixture proportion estimation has been an important ingredient for learning with label noise [31, 19, 36, 5, 39, 13, 12], learning with complementary labels [38], domain adaptation [11, 41], semi-supervised learning [29], anomaly rejection [2, 30], PU learning [8, 23], and multiple instance learning [1], etc. Here, we give a brief summary of the former two applications, and show how the proposed method efficiently solves these problems.

4.1. Learning with label noise

Learning with label noise is a kind of weakly supervised learning where the labels randomly flip from one class to another with some small probabilities. It is often assumed that the flip rates only depend on the class. In this scenario, many works seek ways to design label noise-robust loss functions. For example, [19] viewed the "noisy" and "clean" data as being sampled from two different domains, and then exploited an importance reweighting strategy. [22] and [31] designed cost-sensitive loss functions to mitigate the effects of label noise. The design of noise-robust loss function in these methods often requires the flip rates to be given, which is not true in practice.

The estimation of flip rates remains an open problem. Existing works [31, 19] focused on the anchor set condition. However, both [31] and [19] relied on the conditional probability estimation, which is error-prone when inappropriate training models are chosen. For example, neural networks with too many parameters usually remember all the training examples [40], which leads to incorrect estimation of the proportions. Another problem is that many previous works focus on the binary classification problem and cannot be directly extended into multi-class setting. The proposed non-parametric method exploits kernel mean embedding of distributions to avoid these problem. Even though collecting correct labels for a large-scale dataset is often expensive, it is often assumed that it is easy to obtain labels for some instances [37, 20, 35]. Our proposed method is well suitable for this problem setting.

Here we denote P_ρ as the distribution related to the noisy labels. We observe that the class-conditional distribution $P_{\rho X|\hat{Y}}$ can be a mixture of $P_{X|Y}$, that is,

$$P_\rho(X|\hat{Y}=i) = \sum_{j=1}^{c} P(Y=j|\hat{Y}=i) P(X|Y=j),$$

where Y and $\hat{Y}$ denote the variables of "clean" and "noisy" labels, respectively; c is the class number; $P(Y=j|\hat{Y}=i)$ is the inversed flip rate. The equation holds due to X being conditional independent from $\hat{Y}$ given the correct label Y. According to Theorem 3, to estimate the flip rates, we do not require too many observed examples from the correct distribution. As such, we should be able to estimate the flip rate without having to obtain too much labelled data [37].

4.2. Semi-supervised learning

Labeling a large set of training data is laborious and expensive. In practical applications, we only have access to a small set of labelled examples and a vast quantity of unlabelled examples. Semi-supervised learning aims to extract information from the unlabelled data to guide the learning.

In the traditional semi-supervised learning, many assumptions on the data structure or distribution have been proposed. For example, data from different classes are often assumed to reside in separate manifolds; the class-prior probabilities for the unlabeled data are assumed to be balanced, i.e. $P(Y=i) = 1/c$, or to be similar to those in the labeled examples [42]. However, if these assumptions are violated, the learning process can be biased. [29]

proposed an approach for semi-supervised learning by combining PU learning and NU learning (learning from negative and unlabelled examples). They developed a cost-sensitive loss function without any assumption on the structure or distribution of training data. However, the cost-sensitive loss function relies on the class-prior probabilities on the unlabelled data. To estimate the class prior, we view the distribution of unlabeled data as a mixture as in many other class ratio estimation problems [30, 28], that is,

$$P(X) = \theta_P P(X|Y=+1) + \theta_N P(X|Y=-1),$$

where $\theta_P + \theta_N = 1$. If the distributions $P(X|Y=+1)$ and $P(X|Y=-1)$ are independent, θ_P and θ_N are unique, which can be efficiently obtained by the proposed mixture proportion estimation method.

5. Experiments

In this section, we validate our method by applying it to the aforementioned two applications including learning with label noise and semi-supervised learning. The proposed MPE method under the Linear Independence Assumption is abbreviated as "MPEIA".

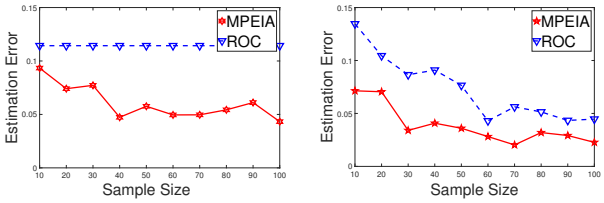
5.1. Learning with label noise

In the label noise setting, we estimate the flip rates using MPE methods. The proposed method is evaluated on both the synthetic and real-world data. The convergence properties and the effects of various flip rates on the proportion estimation are analyzed. Here [31] ("ROC"), [26] ("KM") and [19, 24] ("MCP": Minimum Conditional Probability)[1] are used as the baselines. For fair comparasion, rather than giving only the data from one component as in traditional MPE, KM method is provided with the same number of clean data from each component to estimate each weight. MCP uses the clean data as the data in the anchor set. But the baseline ROC still focuses on the setting that no clean data are provided. The inferior performances do not imply the ineffectiveness of this method.

The MPE methods are tested on both the binary and multi-class settings, and are applied to problems involving both symmetric and asymmetric label noise. Here, we denote Q as the transition matrix, where $Q_{ij} = P(\hat{Y} = j|Y = i)$ is the flip rate from label i to label j. If $Q_{ij}, \forall i \neq j$, is set to the same probability, the label noise is symmetric; otherwise, asymmetric. Even though the proposed method estimates the inversed flip rate $P(Y = j|\hat{Y} = i)$, methods like MCP estimate flip rates. For comparisons, we convert the inversed ones to the flip rates using Bayes' Theorem[2].

[1]The code for KM can be found at http://web.eecs.umich.edu/~cscott/code/kernel_MPE.zip. The code for ROC is from http://web.eecs.umich.edu/~cscott/code/mpe_v2.zip. In this paper, we re-implement the code for MCP.

[2]In the experiments, the class prior of clean data is balanced.



(a) (b)

Figure 2. The convergence analysis of the proposed method. (a) The convergence property w.r.t. the number of examples from all components with the fixed sample size of the mixture. (b) The convergence property w.r.t. the sample sizes of the mixture and components with the fixed ratio $r = 100:1$.

Then the estimation error $\|Q^* - \hat{Q}\|$ is reported, where Q^* is the true transition matrix and $\hat{Q}$ is the estimated transition matrix. In all experiments, the average error of 20 trials is reported for each result.

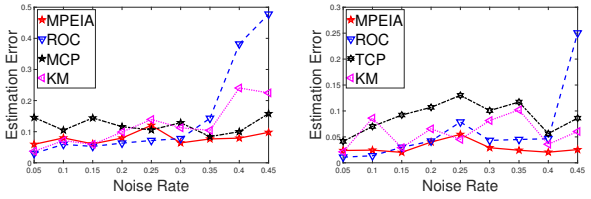
For MMD, the Gaussian kernel is applied with the bandwidth σ being set to the mean value of the pairwise distances of all examples from the mixture and components.

Synthetic data. Here, we use the mixture of two Gaussians to generate the synthetic data: $x \sim \sum_{i=1}^{2} \pi_i \mathcal{N}(\theta_i, \Sigma_i)$, where θ_i are sampled from the uniform distribution $\mathcal{U}(-0.25, 0.25)$, and Σ_i are sampled from the Wishart distribution $0.01 * \mathcal{W}(2 \times \mathbf{I}_2, 7)$. Here data drawn from each Gaussian distribution are in the same class. To generate the label noise data, we symmetrically flip the labels from one class to another with the probability ρ.

In the first experiment, the convergence property of the proposed method is analyzed. The flip rate ρ is set to 0.3. The sample size of the mixture (noisy data) is fixed to 1000, and the number of examples from components varies from 10 to 100. Here the sample size of each component is equal. The result is shown in Figure 2(a). ROC is used as a baseline. We can see that only a relatively small set of clean data is required to estimate the flip rates.

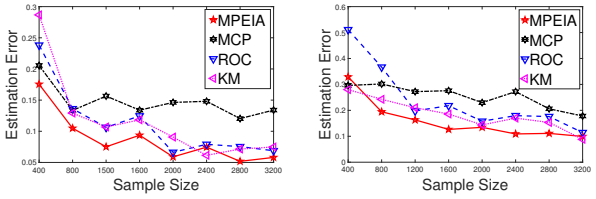
In the rest of this paper, we denote r as the ratio between the sample size of the mixture distribution and the total number of examples from all components. In the next experiment, we fix $r = 100:1$. The number of examples from all components ranges from 10 to 100. The experimental results in Figure 2(b) are also in accordance with the theoretical analysis, that is, the proposed estimator can converge to the true flip rate very quickly.

In the second experiment, we evaluate the performances of the proposed method under different flip rates. r is fixed to 100:1. The flip rate ρ varies from 0.05 to 0.45. Then our method is compared with the state-of-the-art methods. For MCP, the raw data is first mapped to the random Fourier feature space [25] with the dimension of 500. Then we use the support vector machine (SVM) with the linear kernel [4] to estimate the conditional probability $P_\rho(Y|X)$. The hyperparameters are chosen by a 5-fold cross-validation. Figure



(a) (b)

Figure 3. The estimation errors w.r.t. various flip rates. (a) The number of examples from all components is 10. (b) The number of examples from all components is 100.



(a) (b)

Figure 4. The convergence analysis of MPE methods. (a) and (b) show the results on the waveform and spambase datasets, repectively.

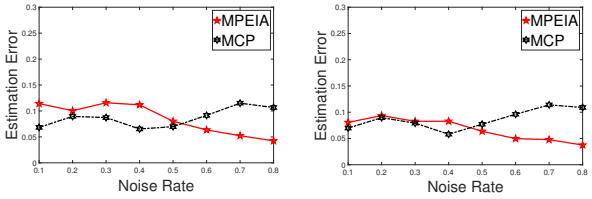
3 shows superiority of the proposed method under most flip rates. Compared to other methods, the proposed method shows more consistent performances for various flip rates.

UCI. We first run our algorithm on binary classification datasets. Two datasets, the "waveform" and "spambase", are taken from the UCI Machine Learning Repository. Here we evaluate the convergence properties of MPEIA, ROC, MCP, and KM. r is fixed to $20:1$. Then we vary the sample size of the mixture from 400 to 3200. The flip rate is fixed to 0.3. Other settings are the same with those in experiments on the synthetic data. The results in Figure 4 show that our method shares the similar convergence properties with KM, and gives the smallest estimation errors.

MNIST. MNIST[3] is a popular handwritten digit database, which consists of 60,000 training examples and 10,000 test examples. The digits range from 0 to 9. For each digit, there are around 6,000 training examples. In this paper, we randomly select a small set of examples from the training set as the clean data, and then randomly flip the labels of the rest training data to generate the mixture.

We evaluate the performances of the proposed method w.r.t. both the symmetric and asymmetric flip rates. Due to the fact that most existing MPE methods, such as ROC and KM, can only be applied to the binary case, the proposed method is compared with MCP. For MPEIA, before estimating the flip rates, we first apply the Principle Component Analysis (PCA) to reduce the feature dimension to 20. For MCP, we follow the method in [24] to estimate the conditional probability. The results for the symmetric label noise

[3]MNIST database is available at http://yann.lecun.com/exdb/mnist/



(a) (b)

Figure 5. The estimation errors w.r.t. various flip rates (MNIST). (a) The ratio between the sample size of noise data and that of clean data is $20:1$. (b) The ratio is $10:1$.

Table 1. The estimation errors of the asymmetric label noise with randomly generated transition matrices (MNIST). The ratio between the sample size of noise data and that of clean data is $20:1$.

MPEIA (ours)	**0.0422**	**0.0448**	**0.0373**	**0.0317**	**0.0477**
MCP	0.1147	0.1519	0.1379	0.1126	0.1442

are shown in Figure 5. We can see that, with such a simple preprocessing, the proposed method can achieve comparable or better performances compared to MCP which needs to train a neural network with two hidden layers. More importantly, Figure 5 shows that the proposed method gives more accurate results than MCP when the flip rates get larger. This is because larger flip rates can adversely affect the conditional probability estimation.

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & .3 & 0 & 0 & 0 & 0 & .7 & 0 & 0 \\ 0 & 0 & 0 & .3 & 0 & 0 & 0 & 0 & .7 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & .3 & .7 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & .7 & .3 & 0 & 0 & 0 \\ 0 & .7 & 0 & 0 & 0 & 0 & 0 & .3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \begin{bmatrix} 1 & \epsilon & \epsilon & \epsilon & \epsilon & \epsilon & \epsilon & \epsilon & \epsilon & \epsilon \\ \epsilon & .99 & \epsilon & \epsilon & .01 & \epsilon & \epsilon & \epsilon & \epsilon & \epsilon \\ \epsilon & \epsilon & .31 & \epsilon & \epsilon & \epsilon & \epsilon & .69 & \epsilon & \epsilon \\ \epsilon & \epsilon & \epsilon & .3 & \epsilon & \epsilon & \epsilon & .01 & .68 & \epsilon \\ \epsilon & \epsilon & \epsilon & \epsilon & 1 & \epsilon & \epsilon & \epsilon & \epsilon & \epsilon \\ \epsilon & \epsilon & \epsilon & \epsilon & \epsilon & .3 & .7 & \epsilon & \epsilon & \epsilon \\ \epsilon & \epsilon & \epsilon & \epsilon & \epsilon & .71 & .29 & \epsilon & \epsilon & \epsilon \\ \epsilon & .7 & \epsilon & \epsilon & \epsilon & \epsilon & \epsilon & .3 & \epsilon & \epsilon \\ \epsilon & \epsilon & \epsilon & \epsilon & .01 & \epsilon & \epsilon & \epsilon & .99 & \epsilon \\ \epsilon & \epsilon & \epsilon & \epsilon & \epsilon & \epsilon & \epsilon & \epsilon & \epsilon & 1 \end{bmatrix}. \quad (4)$$

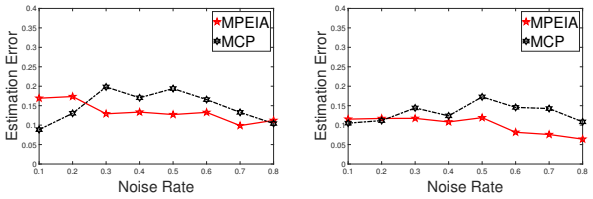
We also give several examples of the asymmetric label noise. First, we construct an example transition matrix with asymmetric flip rates as in [24]. In Eq. (4), the matrix on the left is the ground-truth transition matrix, and the right is the one estimated by the proposed method ($\epsilon \leq 0.005$). Here $r = 20:1$.

Next, we construct the transition matrix with arbitrarily asymmetric flip rates. To get such a transition matrix, a random matrix is first generated, and then each row is normalized to 1. To show the superiority of the proposed method, we repeatedly generate 5 different asymmetric transition matrices, and report the estimation errors in Table 1. We found that, in the case of asymmetric label noise, many examples used to estimate the flip rates are not in the anchor set, which leads to the incorrect estimation. On the contrary, the proposed method performs much better without estimating the conditional probability.

CIFAR10. CIFAR10 is another tiny image dataset [17] with 50,000 training examples and 10,000 test examples. It

Table 2. The estimation errors of the asymmetric label noise with randomly generated transition matrices (CIFAR10). The ratio between the sample size of noise data and that of clean data is $20:1$.

MPEIA (ours)	**0.1298**	**0.1015**	**0.0799**	**0.1378**	**0.1005**
MCP	0.3260	0.2634	0.3287	0.2970	0.3025



(a) (b)

Figure 6. The estimation errors w.r.t. various flip rates (CIFAR10). (a) The ratio between the sample size of noise data and that of clean data is $20:1$. (b) The ratio is $10:1$.

also contains 10 classes of objects. The raw feature of CIFAR10 is far from a good representation. Then in this experiment, we first extract the features of the examples from the pre-trained AlexNet (Caffe Model Zoo [15]) after resizing the images to 227×227. The features of "fc7" layer are used. For MCP, the conditional probability is also estimated by using a neural network with two hidden layers. All other settings are the same with those in the experiments of MNIST. The experimental results in Figure 6 and Table 2 show similar phenomenon with those of MNIST. MCP gives better estimates when the flip rates are very small and symmetric, but much worse estimates when the flip rates get larger or asymmetric. On the contrary, the proposed method provides consistently good estimates in these conditions.

5.2. Semi-supervised learning

In semi-supervised learning, we estimate the class priors using MPE methods. To evaluate the performances, we report the estimates and the corresponding classification errors using PNU (SL) method [29]. The proposed method is compared to PE [6], ED [16] and KM.

The proposed method is also evaluated on the datasets taken from the UCI Machine Learning Repository. Denote all the training examples as $\{x_i\}_{i=1}^n = \mathcal{X}_P \cup \mathcal{X}_N \cup \mathcal{X}_U$, where P, N, U refer to the positive, negative, and unlabelled, respectively. Let $n = n_P + n_N + n_U$, and $\theta_P = P(y = +1)$. To begin with, we re-sample the datasets such that the θ_P of labelled data is set to 0.5, and the θ_P of unlabelled data is 0.2. Then we use MPE methods to estimate the class prior, and apply PNU (SL) method to conduct semi-supervised classification. Other experimental setups, such as the pre-processing of datasets and the kernel used in classification model, follow those in [29].

The results are shown in Table 3. For each dataset, the first row presents the estimate of θ_P for each method, and the second row presents the misclassification rates and standard deviations. For most datasets, MPEIA gives the best estimates of the class prior and best classification results, which verify the effectiveness of the proposed method.

Table 3. The estimates of $\theta_P = 0.2$ of unlabelled data and the misclassification rates of PNU(SL) method. Here $n_P + n_N = 50$, and $n_U = 300$. The average and standard deviation of classification errors over 50 trials for the datasets are reported.

	MPEIA (ours)	ED	KM	PE
Magic	**0.2049**	0.2388	0.2062	0.2676
$d = 10$	**16.4 (1.0)**	17.7 (1.8)	16.9 (1.9)	19.1 (3.9)
SUSY	**0.2040**	0.3992	0.1307	0.1262
$d = 18$	24.4 (2.4)	30.2 (1.2)	**20.2 (0.3)**	20.6 (0.7)
Waveform	**0.1892**	0.3407	0.2977	0.3180
$d = 21$	**9.2 (1.4)**	15.2 (0.9)	14.8 (0.6)	14.9 (1.5)
ijcnn1	0.4509	0.4846	**0.1096**	0.4151
$d = 22$	36.4 (5.4)	36.9 (5.8)	**18.9 (1.1)**	28.7 (5.7)
Spambase	**0.2909**	0.3178	0.4108	0.3830
$d = 57$	14.8 (0.9)	15.1 (1.0)	15.1 (1.9)	**13.3 (2.6)**
a9a	**0.1990**	0.2737	0.2062	0.4488
$d = 83$	**18.1 (0.9)**	18.6 (1.6)	18.1 (1.1)	25.5 (3.2)
w8a	**0.2833**	0.4056	0.3109	0.3666
$d = 300$	**18.9 (2.4)**	21.7 (3.8)	24.8 (1.2)	20.9 (1.9)

6. Conclusion

The MPE problem is addressed in a new setting where the samples from the mixture and all components are given. By exploiting the independence assumption and the kernel mean embedding of distributions, the proposed estimator is ensured to converge to the unique solution at the rate of $O(1/\min(n^{\frac{1}{4}}, n_0^{\frac{1}{4}}))$. The proposed method requires to solve only a simple convex quadratic programming and can be easily applied to some popular applications. The experiments demonstrate that, with only a small number of examples from components, our method is effective to deal with complex label noise and to estimate class priors.

Acknowledgement

This research was supported by Australian Research Council Projects FL-170100117 and DP-180103424. This work was partially supported by SAP SE. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X Pascal GPU used for this research. This research was partially supported by research grant from Pfizer titled "Developing Statistical Method to Jointly Model Genotype and High Dimensional Imaging Endophenotype". We are also grateful for the computational resources provided by Pittsburgh Super Computing grant number TG-ASC170024.

References

[1] J. Amores. Multiple instance classification: Review, taxonomy and comparative study. *Artificial Intelligence*, 201:81–105, 2013. 5

[2] G. Blanchard, G. Lee, and C. Scott. Semi-supervised novelty detection. *Journal of Machine Learning Research*, 11(Nov):2973–3009, 2010. 2, 5

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003. 1

[4] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27, 2011. 6

[5] J. Cheng, T. Liu, K. Ramamohanarao, and D. Tao. Learning with bounded instance-and label-dependent label noise. *arXiv preprint arXiv:1709.03768*, 2017. 5

[6] M. C. Du Plessis and M. Sugiyama. Class prior estimation from positive and unlabeled data. *IEICE Transactions on Information and Systems*, 97(5):1358–1362, 2014. 2, 8

[7] M. C. Du Plessis and M. Sugiyama. Semi-supervised learning of class balance under class-prior change by distribution matching. *Neural Networks*, 50:110–119, 2014. 2

[8] C. Elkan and K. Noto. Learning classifiers from only positive and unlabeled data. In *KDD*, pages 213–220. ACM, 2008. 1, 2, 5

[9] B. Frénay and M. Verleysen. Classification in the presence of label noise: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 25(5):845–869, 2014. 1

[10] K. Fukumizu, A. Gretton, B. Schölkopf, and B. K. Sriperumbudur. Characteristic kernels on groups and semigroups. In *NIPS*, pages 473–480, 2009. 4

[11] M. Gong, K. Zhang, T. Liu, D. Tao, C. Glymour, and B. Schölkopf. Domain adaptation with conditional transferable components. In *ICML*, pages 2839–2848, 2016. 5

[12] B. Han, I. W. Tsang, and L. Chen. On the convergence of a family of robust losses for stochastic gradient descent. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 665–680. Springer, 2016. 5

[13] B. Han, I. W. Tsang, L. Chen, P. Y. Celina, and S.-F. Fung. Progressive stochastic learning for noisy labels. *IEEE Transactions on Neural Networks and Learning Systems*, 2018. 5

[14] A. Iyer, S. Nath, and S. Sarawagi. Maximum mean discrepancy for class ratio estimation: Convergence bounds and kernel selection. In *ICML*, pages 530–538, 2014. 2, 3

[15] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014. 8

[16] H. Kawakubo, M. C. Du Plessis, and M. Sugiyama. Computationally efficient class-prior estimation under class balance change using energy distance. *IEICE Transactions on Information and Systems*, 99(1):176–186, 2016. 8

[17] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. 2009. 7

[18] Y. Li, J. Yang, Y. Song, L. Cao, J. Luo, and J. Li. Learning from noisy labels with distillation. *arXiv preprint arXiv:1703.02391*, 2017. 2

[19] T. Liu and D. Tao. Classification with noisy labels by importance reweighting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(3):447–461, March 2016. 2, 3, 5, 6

[20] I. Misra, C. Lawrence Zitnick, M. Mitchell, and R. Girshick. Seeing through the human reporting bias: Visual classifiers from noisy human-centric labels. In *CVPR*, pages 2930–2939, 2016. 5

[21] K. Muandet, K. Fukumizu, B. Sriperumbudur, B. Schölkopf, et al. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141, 2017. 3

[22] N. Natarajan, I. S. Dhillon, P. K. Ravikumar, and A. Tewari. Learning with noisy labels. In *NIPS*, pages 1196–1204, 2013. 1, 5

[23] G. Niu, M. C. du Plessis, T. Sakai, Y. Ma, and M. Sugiyama. Theoretical comparisons of positive-unlabeled learning against positive-negative learning. In *Advances in Neural Information Processing Systems*, pages 1199–1207, 2016. 1, 5

[24] G. Patrini, A. Rozza, A. Menon, R. Nock, and L. Qu. Making neural networks robust to label noise: A loss correction approach. In *CVPR*, 2017. 1, 2, 6, 7

[25] A. Rahimi and B. Recht. Random features for large-scale kernel machines. In *NIPS*, pages 1177–1184, 2008. 6

[26] H. Ramaswamy, C. Scott, and A. Tewari. Mixture proportion estimation via kernel embeddings of distributions. In *ICML*, pages 2052–2060, 2016. 1, 2, 5, 6

[27] C. E. Rasmussen. The infinite gaussian mixture model. In *Advances in neural information processing systems*, pages 554–560, 2000. 1

[28] S. Sabato and N. Tishby. Multi-instance learning with any hypothesis class. *Journal of Machine Learning Research*, 13(Oct):2999–3039, 2012. 6

[29] T. Sakai, M. C. d. Plessis, G. Niu, and M. Sugiyama. Semi-supervised classification based on classification from positive and unlabeled data. In *ICML*, 2017. 1, 2, 5, 8

[30] T. Sanderson and C. Scott. Class proportion estimation with application to multiclass anomaly rejection. In *AISTATS*, pages 850–858, 2014. 2, 5, 6

[31] C. Scott. A rate of convergence for mixture proportion estimation, with application to learning from noisy labels. In *AISTATS*, 2015. 1, 2, 3, 5, 6

[32] C. Scott, G. Blanchard, and G. Handy. Classification with asymmetric label noise: Consistency and maximal denoising. In *COLT*, pages 489–511, 2013. 2

[33] A. Smola, A. Gretton, L. Song, and B. Schölkopf. A Hilbert space embedding for distributions. In *ALT*, pages 13–31. Springer, 2007. 3

[34] B. K. Sriperumbudur, K. Fukumizu, and G. R. Lanckriet. Universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, 12(Jul):2389–2410, 2011. 4

[35] A. Vahdat. Toward robustness against label noise in training deep discriminative neural networks. In *NIPS*, 2017. 5

[36] R. Wang, T. Liu, and D. Tao. Multiclass learning with partially corrupted labels. *IEEE Transactions on Neural Networks and Learning Systems*, 2017. 5

[37] T. Xiao, T. Xia, Y. Yang, C. Huang, and X. Wang. Learning from massive noisy labeled data for image classification. In *CVPR*, pages 2691–2699, 2015. 2, 5

[38] X. Yu, T. Liu, M. Gong, and D. Tao. Learning with biased complementary labels. *arXiv preprint arXiv:1711.09535*, 2017. 5

[39] X. Yu, T. Liu, M. Gong, K. Zhang, and D. Tao. Transfer learning with label noise. *arXiv preprint arXiv:1707.09724*, 2017. 5

[40] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. In *ICLR*, 2017. 5

[41] K. Zhang, B. Schölkopf, K. Muandet, and Z. Wang. Domain adaptation under target and conditional shift. In *ICML*, pages 819–827, 2013. 5

[42] X. Zhu. Semi-supervised learning tutorial. In *ICML*, pages 1–135, 2007. 5