Privacy-preserving Class Ratio Estimation

Arun Iyer
Indian Institute of Technology Bombay,
Mumbai, India. 400076.
aruniyer@cse.iitb.ac.in

J. Saketha Nath*
Indian Institute of Technology Bombay,
Mumbai, India. 400076.
saketh@cse.iitb.ac.in

Sunita Sarawagi†
Indian Institute of Technology Bombay,
Mumbai, India. 400076.
sunita@iitb.ac.in

ABSTRACT

In this paper we present learning models for the class ratio estimation problem, which takes as input an unlabeled set of instances and predicts the proportions of instances in the set belonging to the different classes. This problem has applications in social and commercial data analysis. Existing models for class-ratio estimation however require instance-level supervision. Whereas in domains like politics, and demography, set-level supervision is more common. We present a new method for directly estimating class-ratios using set-level supervision. Another serious limitation in applying these techniques to sensitive domains like health is data privacy. We propose a novel label privacy-preserving mechanism that is well-suited for supervised class ratio estimation and has guarantees for achieving efficient differential privacy, provided the per-class counts are large enough. We derive learning bounds for the estimation with and without privacy constraints, which lead to important insights for the data-publisher. Extensive empirical evaluation shows that our model is more accurate than existing methods and that the proposed privacy mechanism and learning model are well-suited for each other.

1. INTRODUCTION

In this work we study statistical learning models for estimating the proportion of instances belonging to different classes in a given set of instances. Many real-world applications motivate this problem: a health analyst wants to estimate the proportion of individuals susceptible to a disease in a locality, a political analyst wants to estimate the proportion of votes to different parties, and so on. Recent work [13, 29] shows how to train such models when presented with a set of instances each attached with its correct label. Unfortunately, such instance-labeled data is not easily accessible in domains like health and social analysis. Often labels are available only for groups of instances and not individual instances in the group. For instance in voting, we only have access to aggregate number of votes to each party, and not the vote of any one individual. [23] lists other scenarios where instance-level supervision in not plausible. In other cases (example, in health [11]), the labels are private and not easily accessible to an analyst. Much work [15, 24, 4, 14, 16, 1, 30, 7, 27] exists on creating machine learning models over private data. But most of these assume that learning happens within the trust boundary of private data. We are targeting cases where the training data might have to be aggregated from many private organizations for understanding broader trends. Thus, our goal in this work is to learn a model for estimating class proportions under these two constraints: (i) supervision is in the form of label counts on sets of instances, and (ii) the labels of data is private and model training happens outside the trust boundary of owners of private data.

In the first part of the paper we present a new learning model for estimating class ratios with set-labeled data for supervision. Our model directly estimates class ratios in a given set of instances using the principle of Maximum Mean Discrepancy in a Reproducible Kernel Hilbert Space [10]. We show this paradigm to be significantly more accurate than first building classifiers with set-labeled data using the techniques of [22, 28, 23, 25, 21] and then estimating proportions in predicted instance labels. We theoretically analyze our model and show that our model is statistically consistent. More interestingly, the analysis shows that our model performs best when trained on few sets, each with a large number of examples. In contrast, existing models for training with set-labeled data (e.g., [28]) prefer many, small sets. Since large sets imply easier hiding of private labels, our model is particularly compatible with the goals of learning under privacy constraints. More importantly, the existing models assume that the training and test distributions are the same, which defeats the very purpose of class-ratio estimation. Our empirical results confirm that existing models perform poorly when the class-ratios in the training and test are different; while the proposed models handle such distribution shift well.

Next we extend our model to handle data privacy. In this work we focus only on protecting the privacy of the class labels. This setting has been proposed earlier in [3, 21] and is of interest in domains like health and finance where some fields are more private than others. In this work, we consider the popular (ϵ, δ)-differential privacy [8, 2] as the

*Work done while on sabbatical at Microsoft, Hyderabad, India.
†Work done while on sabbatical at Google, Mountain View, USA.

KDD '16, August 13 - 17, 2016, San Francisco, CA, USA

ISBN 978-1-4503-4232-2/16/08...$15.00

DOI: http://dx.doi.org/10.1145/2939672.2939806

definition of privacy. A widely used mechanism for enforcing differential privacy is based on adding a Laplace noise [8, 20, 18]. We show that this mechanism distorts proportions too much. We propose a new label privacy preserving mechanism that is well-suited for the supervised class-ratio estimation problem. We theoretically analyze our mechanism and show that our method guarantees $\epsilon \to 0, \delta \to 0$ as long as each class has large enough counts, independent of the number of classes. In contrast, existing mechanisms based on Lapace noise have a (non-zero) lower bound on ϵ, which worsens with the number of classes. Empirically also, we show that our mechanism provides much lower distortion particularly for large number of classes.

Our privacy mechanism and learning algorithm are designed to be maximally compatible with each other and aligned to the goals of creating an accurate model while guaranteeing differential privacy. Our privacy mechanism preserves class ratios and incurs low distortion when set sizes are large. Our learning algorithm yields high accuracy when each set is large and the number of sets is small. These together lead to much lower estimation error than can be obtained using existing learning models and existing privacy mechanisms.

In summary, the main contributions of this work are:

1. We designed a model for estimating class ratios that can be trained with set-labeled data. We show that the proposed model outperforms existing ones, especially when the distribution-shift between the training and test sets is high.

2. We theoretically bound the estimation error of our model. The bound shows that our model is statistically consistent, and has low error when the number of sets is minimized (equal to number of classes), each set is large, and label proportions in each training set are skewed. In contrast, existing methods of creating classification models from set-labeled data prefer many, small sets.

3. We propose a new mechanism for achieving (ϵ, δ) differential privacy of a multiset of labels. Unlike existing mechanisms that add independent Laplace noise to each label count, our design preserves class proportions more effectively for increasing number of classes. We analyze our mechanism to show conditions for achieving differential privacy (with $\epsilon \approx 0, \delta \approx 0$).

4. We extend the learning model as well as its analysis for privacy-protected data. We show that the proposed learning and privacy mechanisms are well-suited for each other. In particular we show some conditions that are common for achieving efficiency in both these phases. We also identify that the skew in class-ratios across the sets in training data controls the trade-off between learning and privacy protection.

5. Empirical evaluation of our method on several large real-datasets shows that: (i) our learning model incurs much lower error than baselines designed over existing models trained from set-labeled data [28] and can reduce error by as much as 70% for skewed test sets. (ii) Our privacy mechanism achieves much lower distortion than existing methods, particularly for large class sizes. For example, on a 5-class dataset, our method distorts proportions by only 0.06 whereas the best existing method distorts by 0.15. (iii) We show that the combination of our learning algorithm and privacy mechanism is able to provide significantly more accurate estimates than existing methods. On an Mnist dataset, our model with our mechanism incurs 60% lower error than with Laplace.

The rest of the paper is organized as follows. We first present and analyze our learning model for class ratio estimation with set-labeled data in Section 2. Next in Section 3, we present our mechanism of enforcing differential privacy on label proportions and analyze our learning model with privacy protected data. In Section 4 we empirically compare our learning model and privacy mechanism. We present related work and conclusions in Sections 5 and 6 respectively.

2. SUPERVISED CLASS RATIO ESTIMATION

In this section we define the supervised class ratio estimation problem and discuss models for it.

2.1 Problem statement

We start with a formal definition of the supervised class ratio estimation problem. Let $\mathcal{X}$ be the set of all instances and $\mathcal{Y} = \{1, \ldots, c\}$ be the set of all labels. Our goal is to design an estimator $\mathcal{M}$ that for any given set $U(\subset \mathcal{X})$ estimates the true probabilities of the various classes: $\rho_u = [\rho_{u1}, \ldots, \rho_{uc}]^\top$ in the distribution from which U was sampled. In this paper we use class-ratios or class-proportions to refer to such ρ. Since U is a finite sample, in practice we will only be able to estimate the sample proportions $\hat{\rho}_u = [\hat{\rho}_{u1}, \ldots, \hat{\rho}_{uc}]^\top$ that denotes the fraction of instances in U belonging to the various classes.

To facilitate the estimation, supervised training data $\mathcal{D}$ consisting of M sets of instances, $\mathcal{S}_i \subset \mathcal{X},\ i = 1, \ldots, M$, and the corresponding fractions of instances belong to the various classes, $\hat{\rho}_i$ are provided, i.e., $\mathcal{D} = \{(\mathcal{S}_i, \hat{\rho}_i) : i = 1 \ldots M\}$. We will use n_i to denote the number of instances in set $\mathcal{S}_i$. We call such $\mathcal{D}$ set-labeled data. Such set-labeled supervision is much weaker than in standard classification where labels are associated with each instance in the training data[1].

Needless to say, we need to assume that the affine-hull of the label proportions, $\hat{\rho}_i\ \forall i = 1, \ldots, M$, contains the c-dimensional simplex[2]:

$$\Delta_c \subset \text{Aff}(\{\hat{\rho}_i\ \forall i\}) \qquad (\mathcal{A}0)$$

If this is not the case, then it is easy to see that some classes may go totally un-represented.

A standard assumption in supervised learning is that the joint-distribution over $\mathcal{X} \times \mathcal{Y}$ is unchanged between training and test data. In contrast, in this setting we allow the distribution of the class labels to differ among the sets in training and test. For instance, the class proportions ρ_u in the unlabeled set U can be arbitrary and be very different

[1]However this setting still falls under standard supervised learning as the prediction/estimation is also over sets and not over instances

[2]The affine hull of a set of vectors V, denoted by Aff(V), is the set of all linear combinations of vectors in V such that the combining coefficients sum to unity.

from those in the training set. Our only distributional assumption about the training and test sets is that the class conditional distribution $P(\mathbf{x}|y)$ is unchanged. That is, for any two sets $\mathcal{S}', \mathcal{S}$ in the training and/or test set

$$P(\mathbf{x}|y, S) = P(\mathbf{x}|y, S') \quad \forall y \in \mathcal{Y}, \tag{$\mathcal{A}1$}$$

Note this is a much weaker assumption than in classification and in existing models for class ratio estimation (e.g., [23, 28]) which assume that both $P(\mathbf{x}|y)$ and $P(y)$ are preserved in the training and test sets. Requiring the $P(y)$ distribution to be unchanged defeats the very goal of estimating class proportions.

2.2 Existing models

One method to solve the above problem using existing literature, is to tap into the recent work on learning classifiers from set-labeled training data [22, 28, 23, 25]. The goals of all these work is to create a classifier to predict labels of individual instances given weak supervision in the form of set-labeled data during training. Using such a classifier, we can estimate label proportions in any set U as follows: for each $\mathbf{x}_j \in U$, invoke classifier to get predicted label $\hat{y_j}$ and then estimate/approximate the class proportions as the fractions of instances belonging to the various classes.

Firstly, such methods assume that the training and test distributions are the same and hence will perform poorly. Also, recent work [29, 13] shows that learning methods that directly estimate the class-ratios out-perform such per-instance predictive models. The models of [29, 13] require instance-labeled data and cannot be trained with set-labeled training data, limiting their pragmatic applicability [23]. We are aware of no other work that proposes a method of direct class ratio estimation with set-labeled data. In the next section we present the first such model.

2.3 Our Model

In this section we detail the proposed class ratio estimation algorithm.

We begin by recalling the basic assumption $\mathcal{A}0$, which motivates us to parameterize the class ratios in the unlabeled set using:

$$\rho_u = \sum_{i=1}^{M} \alpha_i \rho_i \tag{1}$$

Then the problem of class ratio estimation boils down to that of estimating the parameters α_i.

We now make the following identifiability assumption without which the class ratio estimation problem is undefined and no algorithm can identify the true class ratios:

$$\boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2 \Rightarrow \sum_{y \in \mathcal{Y}} P(\mathbf{x}|y)\theta_{1y} \neq \sum_{y \in \mathcal{Y}} P(\mathbf{x}|y)\theta_{2y}, \tag{$\mathcal{A}2$}$$

in other words, we are assuming that the class conditionals are linearly independent. Let us denote the class conditional distribution that is common for all sets by $P(\mathbf{x}|y)$.

With the above in place, we note that:

$$\rho_u = \sum_{i=1}^{M} \alpha_i \rho_i \iff \sum_{y=1}^{c} P(\mathbf{x}|y)\rho_{uy} = \sum_{y=1}^{c} P(\mathbf{x}|y) \sum_{i=1}^{M} \alpha_i \rho_{iy}$$

$$\iff P(\mathbf{x}, U) = \sum_{i=1}^{M} \alpha_i P(\mathbf{x}, \mathcal{S}_i),$$

where $P(\mathbf{x}, \mathcal{S}_i)$ denotes the marginal distribution over $\mathcal{X}$ that generated $\mathcal{S}_i$.

Thus, to solve for α we need a method to represent and compare the distributions of each set. Recently, a powerful tool for such algebraic operations on distributions has been provided by the concept of Maximum Mean Discrepency (MMD) on a Reproducible Kernel Hilbert Space (RKHS) [10]. Examples of some algorithms based on MMD are: handling covariance shift [10], the two-sample problem [9], class ratio estimation [29, 13] and training deep generative neural networks [19]. MMD uses the notion of embedding distributions in the RKHS of a kernel. Using this we find α's by minimizing the distance between $P(\mathbf{x}, U)$ and $\sum_{i=1}^{M} \alpha_i P(\mathbf{x}, \mathcal{S}_i)$.

Accordingly, we define K to be a characteristic kernel and let $\mathcal{H}$ denote the RKHS induced by K. Let $\Phi : \mathcal{X} \mapsto \mathcal{H}$ denote the canonical feature map induced by the kernel onto the RKHS. Let

$$\bar{\Phi}_i(\mathbf{x}) = \mathbb{E}_{P(\mathbf{x},S_i)}\Phi(\mathbf{x}), \forall i \in [1, \ldots, M], \tag{2}$$

and

$$\bar{\Phi}_U(\mathbf{x}) = \mathbb{E}_{P(\mathbf{x},U)}\Phi(\mathbf{x}), \tag{3}$$

where $\mathbb{E}_{P(\mathbf{x},\mathcal{S})}h(\mathbf{x})$ denotes the expectation of $h(\mathbf{x})$ taken under the distribution $P(\mathbf{x}, \mathcal{S})$.

Our objective of finding α's such that $\sum_{i=1}^{M} \alpha_i P(\mathbf{x}, S_i) = P(\mathbf{x}, U)$ can now be posed as the following optimization problem:

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^M} \left\| \sum_{i=1}^{M} \alpha_i \bar{\Phi}_i(\mathbf{x}) - \bar{\Phi}_U(\mathbf{x}) \right\|_{\mathcal{H}} \tag{4}$$

The class-ratios can then be computed using $\rho_u = \mathcal{P}\alpha^*$, where $\mathcal{P}$ is the matrix with i^{th} column as ρ_i and α^* is the optimal solution of (4). Now since $\bar{\Phi}_i, \bar{\Phi}_U$ as well as $\mathcal{P}$ are not available, we instead approximate them using the following empirical estimates:

$$\bar{\Phi}_i \approx \hat{\Phi}_i \equiv \frac{1}{n_i} \sum_{\mathbf{x} \in S_i} \Phi(\mathbf{x})$$

and

$$\bar{\Phi}_U \approx \hat{\Phi}_U \equiv \frac{1}{n_u} \sum_{\mathbf{x} \in U} \Phi(\mathbf{x})$$

where $n_i = |S_i|$ and $n_u = |U|$. And,

$$\mathcal{P} \approx \hat{\mathcal{P}} \equiv [\hat{\rho}_1 \ldots \hat{\rho}_M]$$

This leads to the following estimate for α's:

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^M} \left\| \sum_{i=1}^{M} \alpha_i \hat{\Phi}_i - \hat{\Phi}_U \right\|_{\mathcal{H}} \tag{5}$$

and the following for the class-ratios: $\hat{\mathcal{P}}\hat{\alpha}$, where $\hat{\alpha}$ is the optimal solution of (5).

Now, because of the approximations it may as well happen that $\hat{\mathcal{P}}\hat{\alpha} \notin \Delta_c$. Hence we finally propose to employ the following estimate: $\hat{\rho}_u \equiv Proj_{\Delta_c}\left(\hat{\mathcal{P}}\hat{\alpha}\right)$, where $Proj_V(v)$ denotes the projection of vector v onto the set V.

The analysis presented in 2.4 shows that the above two-step approximation indeed leads to a statistically consistent alogrithm. More importantly, it shows that: i) $M = c$ number of sets, each with large number of examples leads to efficient estimation. Note that this is completely in contrast

with the $\propto$-SVM [28]. This feature of our algorithm, as we shall see later, naturally leads to high-levels of label privacy; thus achieving good trade-off between estimation accuracy and label privacy, and, ii) the more the label purity[3] in the sets, the better the estimation.

2.4 Main results from the theoretical analysis

In this section we present the learning bounds associated with the proposed algorithm. We begin by rewriting (4) and (5) respectively as:

$$\min_{\boldsymbol{\alpha}} \left\| \bar{A}\boldsymbol{\alpha} - \bar{a} \right\| \tag{6}$$

where $\bar{A} = [\bar{\Phi}_1(\mathbf{x}), \ldots, \bar{\Phi}_M(\mathbf{x})]$ and $\bar{a} = [\bar{\Phi}_U(\mathbf{x})]$ and

$$\min_{\boldsymbol{\alpha}} \left\| \hat{A}\boldsymbol{\alpha} - \hat{a} \right\| \tag{7}$$

where $\hat{A} \equiv [\hat{\Phi}_1, \ldots, \hat{\Phi}_M]$ and $\hat{a} = [\hat{\Phi}_U]$.

For the analysis we make another very mild assumption:

$$\hat{A} \text{ has full column rank.} \qquad (\mathcal{A}4)$$

Note that this is satisfied whenever the instances in $\mathcal{D}_x$ are unique, since the kernel is universal. To simplify the notation in the learning bound we assume that the kernel is normalized and hence $\max_{x \in \mathcal{X}} \|\Phi(x)\| = 1$.

THEOREM 1. *Given the notation, and assumptions stated above, the error of our estimated proportions $\hat{\rho}_u$ from the true proportions ρ_u is upper bounded with at least probability $1-\delta$ as follows:*

$$\|\hat{\rho}_u - \rho_u\| \leq \frac{\sqrt{\sum_{i=1}^{M} \mathcal{C}_\delta(n_i)^2}}{minsing\left(\bar{A}\right)} \left(1 + Q(\mathcal{D})\right) + Q(\mathcal{D})\, \mathcal{C}_\delta(n_u)$$

where $\mathcal{C}_\delta(n)$ is a confidence term defined by $\mathcal{C}_\delta(n) \equiv \frac{2}{\sqrt{n}} + \sqrt{\frac{\ln(2(2M+1)/\delta)}{2n}}$, $minsing\left(\bar{A}\right)$ is the minimum singular value of $\bar{A}$ (maxsing is analogously defined), and $Q(\mathcal{D})$ is, what we call the condition number of the training set, defined by $Q(\mathcal{D}) \equiv \frac{maxsing\left(\hat{\mathcal{P}}\right)}{minsing\left(\hat{A}\right)}$.

We detail in Section 2.4.1 the valuable insights provided by the bound and then in Section 2.4.2 present a sketch of the proof.

2.4.1 Properties of our estimator

First, the confidence term decays at $O(\frac{1}{\sqrt{n}})$ with growing sizes of the sets, and shows that the estimation error reduces with increasing number of points in both the training and test sets. The model prefers smaller number of sets with large number of points in each set[4]. Note that this is in contrast with $\propto$-SVM and is an attractive feature of our algorithm wrt. privacy.

Second, we notice that the upper bound vanishes to zero as the training set sizes n_i and test set size n_u approach infinity. This indicates that our estimator is consistent.

[3]Purity of set increases as the proportion of instances with the majority label approaches unity.

[4]However, since we want to span the entire convex hull spanned by the class conditional distributions, the minimum number of sets we must have is c.

The condition number $Q(\mathcal{D})$ brings out more desirable properties of our estimator. This term simply measures how "close-by" the sets are with respect to the y distribution (via the term maxsing $\left(\hat{\mathcal{P}}\right)$ in the numerator) and the $\mathbf{x}$ distribution (via the term minsing $\left(\hat{A}\right)$ in the denominator). The more diverse the class-ratios are across the sets, less is the condition number and vice-versa. It is easy to see that the condition number is the highest, reaching ∞, when the class-ratios are almost the same across the sets (because the minsing will then be near zero). And is the least, approaching unity, when the class-ratios are orthogonal to each other. This indicates that the estimator performs best when the class proportions in the training set are skewed towards any one class.

2.4.2 Proof of Theorem 1

To prove the theorem, we begin with the observation that[5]:

$$\begin{aligned}
\|\hat{\rho}_u - \rho\| &= \|Proj_{\Delta_c}\left(\hat{\mathcal{P}}\hat{\boldsymbol{\alpha}}\right) - \mathcal{P}\boldsymbol{\alpha}^*\| \\
&\leq \|\hat{\mathcal{P}}\hat{\boldsymbol{\alpha}} - \mathcal{P}\boldsymbol{\alpha}^*\| \\
&\leq \|\boldsymbol{\alpha}^*\|\|\hat{\mathcal{P}} - \mathcal{P}\| + \|\hat{\mathcal{P}}\|\|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\| \\
&\leq \|\boldsymbol{\alpha}^*\|\|\hat{\mathcal{P}} - \mathcal{P}\|_F + \|\hat{\mathcal{P}}\|\|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\| \qquad (8)
\end{aligned}$$

We now proceed to bound the two difference terms above. The first difference term accounts for the error that may arise from the errors in the estimates of the class proportions of each set calculated from finite sets. Lemma 1 gives the error bound for this difference term. The proof for the bound for the second difference term proceeds in two steps. In the first step via Lemma 2, we bound the difference term with the difference in the objective function of the respective optimization problems in Equations 6 and 7. In the second step in Lemma 3, we bound the difference in the objective functions in terms of known quantities.

LEMMA 1. *With probability $1 - \delta\frac{M}{2M+1}$*

$$\|\hat{\mathcal{P}} - \mathcal{P}\|_F \leq \sqrt{\sum_{i=1}^{M} \left(\frac{2}{\sqrt{|S_i|}} + \sqrt{\frac{\ln(2(2M+1)/\delta)}{2|S_i|}} \right)^2}$$

PROOF. Notice that, $\|\hat{\mathcal{P}} - \mathcal{P}\|_F = \sqrt{\sum_{i=1}^{M} \|\hat{\rho}_i - \rho_i\|^2}$ where ρ_i is the true class proportion in the set S_i. We can now proceed to bound each term of this summation using Theorem 27 from [17]. □

LEMMA 2. $\|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\| \leq \sqrt{\frac{\|\hat{A}\boldsymbol{\alpha}^* - \hat{a}\|^2 - \|\hat{A}\hat{\boldsymbol{\alpha}} - \hat{a}\|^2}{minsing\left(\hat{A}\right)}}$

PROOF. The proof for the above lemma is identical to proof of Lemma 1 in [13]. □

LEMMA 3. *With probability $1 - \delta\frac{M+1}{2M+1}$,*

$$\begin{aligned}
&\sqrt{\|\hat{A}\boldsymbol{\alpha}^* - \hat{a}\|^2 - \|\hat{A}\hat{\boldsymbol{\alpha}} - \hat{a}\|^2} \\
&\leq \|\boldsymbol{\alpha}^*\| \sqrt{\sum_{i=1}^{M} \left(\frac{2}{\sqrt{n_i}} + \sqrt{\frac{\ln(2(2M+1)/\delta)}{2n_i}} \right)^2} \\
&\quad + \frac{2}{\sqrt{n_u}} + \sqrt{\frac{\ln(2(2M+1)/\delta)}{2n_u}}
\end{aligned}$$

[5]Here, the matrix norm is the maximum singular value and the Frobenius norm is highlighted with an 'F' subscript.

PROOF. Note that, $\|\hat{A}\boldsymbol{\alpha}^* - \hat{a}\|^2 - \|\hat{A}\hat{\boldsymbol{\alpha}} - \hat{a}\|^2 \leq \|\hat{A}\boldsymbol{\alpha}^* - \hat{a}\|^2$. From our discussion in Section 2.3, we know that $\sum_{i=1}^{M} \alpha_i^* \bar{\Phi}_i(\mathbf{x}) = \bar{\Phi}_U$. With this knowledge, we get,

$$\|\hat{A}\boldsymbol{\alpha}^* - \hat{a}\| \leq \|\boldsymbol{\alpha}^*\| \sqrt{\sum_{i=1}^{M} \|\hat{\Phi}_i(\mathbf{x}) - \bar{\Phi}_i(\mathbf{x})\|^2} + \left\|\bar{\Phi}_U(\mathbf{x}) - \hat{\Phi}_U(x)\right\|$$

Applying Theorem 27 from [17] to each individual difference term gives us the result. □

Combining results from Lemma 1, 2, and 3 and the fact that $\|\alpha^*\| \leq \frac{1}{\text{minsing}(\bar{A})}$ proves the theorem.

3. PRESERVING LABEL PRIVACY

Let $\mathcal{S} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\}$ represent a typical set in the un-published training data. Now, let us denote the set consisting of only the input instances from $\mathcal{S}$ i.e., $\{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ by $\mathcal{S}^x$. Similary, define $\mathcal{S}^y = \{y_1, \ldots, y_m\}$. Let η and ρ denote the vectors with entries as the number and proportions respectively of the instances belonging to the various classes in $\mathcal{S}$. Note that $\sum_{i=1}^{c} \eta_i = m$ and thus each $\rho_i = \frac{\eta_i}{m}$.

In this section we present a label privacy preserving mechanism g, which takes as input $\mathcal{S}^y$ and outputs a vector $\hat{\theta}$ that is a proxy for ρ, the proportions of instances belonging to the various classes in $\mathcal{S}$. Our goal is for g to satisfy the standard differential privacy requirements [8] which states that the output of g should be randomized such that neighboring data sets assign similar probabilities to any output θ. More formally, let $\mathcal{S}^y_{(i,j)}$ denote the dataset obtained from $\mathcal{S}^y$ by changing the label of an instance from class i to class j. Then g is said to satisfy (ϵ, δ)-differential privacy (DP) iff:

$$P\left[g(\mathcal{S}^y) \in B\right] \leq \delta + e^{\epsilon} P\left[g(\mathcal{S}^y_{(i,j)}) \in B\right],\ \forall\ B \subset \Delta_c,\ \forall\ i \neq j.$$

We seek a mechanism that achieves this privacy while minimally distorting the true ρ-s so that they continue to be useful for the learning algorithm. A popular recipe for (ϵ, δ)-differential privacy is to add to each η_i an independent random noise z_i generated from a Laplace or Gaussian distribution[8]. The resultant $\tilde{\eta}_i = \eta_i + z_i$ may not be positive or sum to m. Recently, [18] proposed a constrained least square step that can be used to convert such $\tilde{\eta}_i$ to a valid output in a post-processing step. Another option is to use the technique of [20] that uses $\tilde{\eta}_i$ to define the parameters of a Dirichlet distribution and sample a valid proportion θ from this distribution. While all these mechanisms achieve (ϵ, δ) privacy, they do so at the cost of a large distortion to the true proportions as we show in Section 4. In this paper we propose a new mechanism that achieves (ϵ, δ) privacy with significantly smaller distortion, provided no class is under-represented. Our mechanism achieves differential privacy with $\epsilon \to 0, \delta \to 0$ asymptotically[6]. In contrast, in the mechanism of [20] ϵ is bounded from below by a value that grows with k.

The key point of our mechanism is to sample θ from a Dirichlet distribution given by:

$$g(\mathcal{S}^y) \sim Dir(\sigma\eta_1, \sigma\eta_2, \ldots, \sigma\eta_k),$$

[6]As $\eta_i \to \infty\ \forall\ i = 1, \ldots, k$.

Here σ is an input-dependent parameter that is tuned to achieve the desired trade-off between the level of privacy and accuracy of the output. We next show how to choose σ for a given (ϵ, δ) and $\mathcal{S}^y$.

Let $f_0 \equiv Dir\left(\sigma\eta_1, \sigma\eta_2, \ldots, \sigma\eta_i, \ldots, \sigma\eta_j, \ldots, \sigma\eta_k\right)$ and $f_{ij} \equiv Dir\left(\sigma\eta_1, \sigma\eta_2, \ldots, \sigma(\eta_i - 1), \ldots, \sigma(\eta_j + 1), \ldots, \sigma\eta_k\right)$ denote the probability density functions of $g(\mathcal{S}^y)$ and $g(\mathcal{S}^y_{(i,j)})$ respectively. The key steps in the analysis are:

1. Finding Θ_{ij} such that $f_0(\theta) \leq e^{\epsilon} f_{ij}(\theta)\ \forall\ \theta \in \Theta_{ij}$. Hence for $B \subset \Theta_{ij}$, we have that differential privacy is satisfied with $\delta = 0$.
2. For $B \not\subset \Theta_{ij}$, it is easy to see that (ϵ, δ)-differential privacy holds if $\delta \geq P\left[g(\mathcal{S}) \notin \Theta_{ij}\right]\ \forall\ i, j$. Hence we need to compute $\max_{i,j} P\left[g(\mathcal{S}) \notin \Theta_{ij}\right]$.

Accordingly, let us first find $\Theta_{i,j}$: Now, $f_0(\theta) \leq e^{\epsilon} f_{ij}(\theta)$

$$\iff \quad \frac{\theta_i^{\sigma}}{\theta_j^{\sigma}} \frac{\Gamma(\sigma\eta_i - \sigma)}{\Gamma(\sigma\eta_i)} \frac{\Gamma\left(\sigma\eta_j + \sigma\right)}{\Gamma\left(\sigma\eta_j\right)} \leq e^{\epsilon},$$

$$\iff \quad \frac{\theta_i}{\theta_j} \leq \Lambda_{ij}^{\frac{1}{\sigma}} e^{\frac{\epsilon}{\sigma}}$$

where $\Lambda_{ij} \equiv \frac{\Gamma(\sigma\eta_i)}{\Gamma(\sigma\eta_i - \sigma)} \frac{\Gamma\left(\sigma\eta_j\right)}{\Gamma\left(\sigma\eta_j + \sigma\right)}$. Thus,

$$\Theta_{ij} = \left\{\theta \in \Delta_k \mid \frac{\theta_i}{\theta_j} \leq \Lambda_{ij}^{\frac{1}{\sigma}} e^{\frac{\epsilon}{\sigma}}\right\}.$$

Now, for bounding δ, we calculate

$$\delta = P\left[g(\mathcal{S}) \notin \Theta_{ij}\right] = P\left[\theta_i - \Lambda_{ij}^{\frac{1}{\sigma}} e^{\frac{\epsilon}{\sigma}} \theta_j > 0\right] \tag{9}$$

$$= \int_{\Theta_{ij}} Dir\left(\sigma\eta_i, \sigma\eta_j, \sigma(m - \eta_i - \eta_j)\right)$$

The above integral is not easy to solve in closed form but we can use the Chebyshev's inequality to bound it as follows.

Let Z_{ij} denote the random variable $\theta_i - \Lambda_{ij}^{\frac{1}{\sigma}} e^{\frac{\epsilon}{\sigma}} \theta_j$. The expectation of Z_{ij} is $\mathbb{E}\left[Z_{ij}\right] = \hat{\rho}_i - \Lambda_{ij}^{\frac{1}{\sigma}} e^{\frac{\epsilon}{\sigma}} \hat{\rho}_j$ and its variance is given by

$$var\left(\theta_i\right) + \Lambda_{ij}^{\frac{2}{\sigma}} e^{2\frac{\epsilon}{\sigma}} var\left(\theta_j\right) - 2\Lambda_{ij}^{\frac{1}{\sigma}} e^{\frac{\epsilon}{\sigma}} cov\left(\theta_i, \theta_j\right),$$

$$= \frac{\hat{\rho}_i\left(1 - \hat{\rho}_i\right) + \Lambda_{ij}^{\frac{2}{\sigma}} e^{2\frac{\epsilon}{\sigma}} \hat{\rho}_j\left(1 - \hat{\rho}_j\right) + 2\Lambda_{ij}^{\frac{1}{\sigma}} e^{\frac{\epsilon}{\sigma}} \hat{\rho}_i \hat{\rho}_j}{(\sigma m + 1)}$$

Using the Chebyshev's inequality, we obtain δ as:

$$\max_{i \neq j} \frac{\hat{\rho}_i\left(1 - \hat{\rho}_i\right) + \Lambda_{ij}^{\frac{2}{\sigma}} e^{2\frac{\epsilon}{\sigma}} \hat{\rho}_j\left(1 - \hat{\rho}_j\right) + 2\Lambda_{ij}^{\frac{1}{\sigma}} e^{\frac{\epsilon}{\sigma}} \hat{\rho}_i \hat{\rho}_j}{(\sigma m + 1)\left(\hat{\rho}_i - \Lambda_{ij}^{\frac{1}{\sigma}} e^{\frac{\epsilon}{\sigma}} \hat{\rho}_j\right)^2} \tag{10}$$

For a given ϵ, we perform line search over σ to find the largest σ for which we achieve a target δ. We may find no σ for a given δ when any of the η_i-s is too small, in which case one can fallback to any standard mechanism like [20]. In our experiments we performed the search exactly via Equation 9 but for faster search the approximation in Equation 10 might be more useful.

level (ϵ, δ) for given counts η. When any η_i is too small, not all choices of ϵ, δ, may yield a valid σ. The algorithm detects such cases and returns an error. When all η_i-s are large, our mechanism guarantees that differential privacy with arbitrarily small ϵ, and δ is possible as we show in the asympototic analysis below.

In the asymptotic case i.e., as each $\eta_i \to \infty$, $\Lambda_{ij}^{\frac{1}{\sigma}}$ behaves like $\frac{\hat{\rho}_i}{\hat{\rho}_j}$. Using this substitution we obtain that

$$\delta = \frac{\left(\max_{ij} \frac{1-\hat{\rho}_i}{\hat{\rho}_i} + e^{2\frac{\epsilon}{\sigma}} \frac{1-\hat{\rho}_j}{\hat{\rho}_j}\right) + 2e^{\frac{\epsilon}{\sigma}}}{(\sigma m + 1)\left(1 - e^{\frac{\epsilon}{\sigma}}\right)^2}. \quad (11)$$

From the above it is clear that in the asymptotic case, differential privacy with $\epsilon \to 0, \delta \to 0$ is possible as long as the parameter $\sigma \to 0$ such that $\sigma\eta_i \to \infty \ \forall \ i$ and $\frac{\epsilon}{\sigma} \to \kappa$, for some constant $\kappa > 0$. Also, the above analysis shows that sets with near uniform distribution of labels lead to lower δ, and hence better privacy. However, more interestingly, if σ is chosen properly, low values of δ are achievable even with sets with skewed class-ratios (which lead to better class-ratio estimation).

In Section 4 we show that even in the finite set case our mechanism provides much smaller error for a given (ϵ, δ) requirement than existing methods.

3.1 Estimation under Privacy Constraints

In this section we extend the proposed class-ratio estimation algorithm as well as its analysis for the case where the training set is perturbed using the privacy preserving mechanism proposed in the previous section. Let $\sigma_i(\epsilon, \delta)$ be the Dirichlet mechanism parameter set to achieve (ϵ, δ)-differential privacy over the set $\mathcal{S}_i$.

The estimation algorithm is now given by: $\tilde{\rho}_u \equiv Proj_{\Delta_c}\left(\tilde{\mathcal{P}}\hat{\alpha}\right)$, where $\tilde{\mathcal{P}}$ is the matrix with i^{th} column as $g(\mathcal{S}_i^y)$. In other words, the algorithm simply uses the output of the Dirichlet mechanism as a proxy for the fraction of labels in the set. It is easy to see that the only modification in the analysis is that the $\|\hat{\mathcal{P}} - \mathcal{P}\|_F$ term in (8) is now replaced by $\|\tilde{\mathcal{P}} - \mathcal{P}\|_F$. Again, $\|\tilde{\mathcal{P}} - \mathcal{P}\|_F \leq \|\hat{\mathcal{P}} - \mathcal{P}\|_F + \|\tilde{\mathcal{P}} - \hat{\mathcal{P}}\|_F$. Hence, in the following we analyze only the new term $\|\tilde{\mathcal{P}} - \hat{\mathcal{P}}\|_F$ and leave details of the other analogous terms to the reader. More importantly, analysis of this term illustrates the suitability of the proposed Dirichlet mechanism for estimating class-ratios using the proposed MMD based algorithm. Towards this goal we present the following lemma:

LEMMA 4. *With probability $1 - \zeta$ we have:*

$$\|\tilde{\mathcal{P}} - \hat{\mathcal{P}}\|_F \leq \frac{Mc}{\sqrt{\zeta}} \max_{i \in \{1,\ldots,M\}; y \in \mathcal{Y}} \sqrt{\frac{\hat{\rho}_{iy}(1 - \hat{\rho}_{iy})}{\sigma_i(\epsilon, \delta) m_i + 1}},$$

where all notations are as defined in Section 2.

Before we present the proof, it is insightful to note the following:

- The best case for privacy is when $\epsilon \to 0, \delta \to 0$. From the discussions in section 3, it is known that this is plausible if $\sigma_i(\epsilon, \delta) m_i \to \infty$.
- Interestingly, the bound in the above lemma also converges to zero as $\sigma_i(\epsilon, \delta) m_i \to \infty$. Hence the proposed Dirichlet mechanism is well-suited for class-ratio estimation using the MMD based algorithm.

PROOF. We begin by noting that:

$$P\left[\|\tilde{\mathcal{P}} - \hat{\mathcal{P}}\| > \tau\right] \leq \sum_{i=1}^{M}\sum_{y=1}^{c} P\left[\|\tilde{\rho}_{iy} - \hat{\rho}_{iy}\|^2 > \frac{\tau^2}{Mc}\right]$$

Also,

$$P\left[\|\tilde{\rho}_{iy} - \hat{\rho}_{iy}\|^2 > \frac{\tau^2}{Mc}\right] \leq \frac{Mc\ var(\tilde{\rho}_{iy})}{\tau^2} = \frac{Mc\ \hat{\rho}_{iy}(1 - \hat{\rho}_{iy})}{\tau^2(\sigma_i(\epsilon, \delta) m_i + 1)}$$

The result follows by choosing the RHS of the last inequality to be $\frac{\zeta}{Mc}$. □

We conclude this section by noting that the insights provided from the learning bounds are indeed very useful for the data-publisher, who may have access to instance-level supervised data. The bounds suggest that the data-publisher must create c sets, with almost equal number of instances in each set, and such that each set has moderate level of skew in the class-proportions (while ensuring $\mathcal{A}0$ is satisfied). As the skew increases, the learning becomes more efficient but privacy might suffer. Thus, the data publisher could prefer the largest skew that satisfies his privacy requirements possibly as per guidelines laid in [12].

4. EXPERIMENTS

In this section we first show that our mechanism for enforcing (ϵ, δ) privacy on a set of labels induces less distortion than existing methods (Section 4.1). We next show that our learning model for estimating class proportions is more accurate than existing models for both undistorted training data (Section 4.2) and privacy protected data (Section 4.3)

4.1 Privacy preserving mechanisms

In this section we compare different privacy preserving mechanisms for publishing a set of labels as discussed in Section 3. For these experiments, the input is a set of c counts $\eta = (\eta_1, \ldots, \eta_c)$ and we generate perturbed $\hat{\eta}$ via different mechanisms of (ϵ, δ) differential privacy. We measure distortion as the L1 distance between the true and perturbed counts scaled by the size of the set.

We compare the following three methods

1. **Laplace:** In this method we first add Laplace noise [8] to each η_i value using a sensitivity parameter 2. We then post process using constrained least square to make perturbed counts be non-negative and sum to m as described in [18].
2. **Gaussian:** Same as above but with the noise generated via a Gaussian distribution instead of Laplace as described in [8, 2].
3. **Laplace Prior:** This is the mechanism proposed in [20] where the counts distorted with Laplace noise are used to define parameters of a Dirichlet distribution. The output proportions are sampled from this Dirichlet.
4. **Scaled Dirichlet:** This is our mechanism described in Section 3.

For these experiments we set the default value of c to 5, set size (m) to 1000, class proportion for first $c - 1$ classes to a $\rho_1 = 0.05$ and the last class to $1 - (c - 1)\rho_1$, and (ϵ, δ) to (0.05, 0.05). All reported numbers are averaged over twenty random samples of perturbed outputs for the same input. We study distortion between the true and perturbed proportions under varying values of the default parameters in Figure 1.

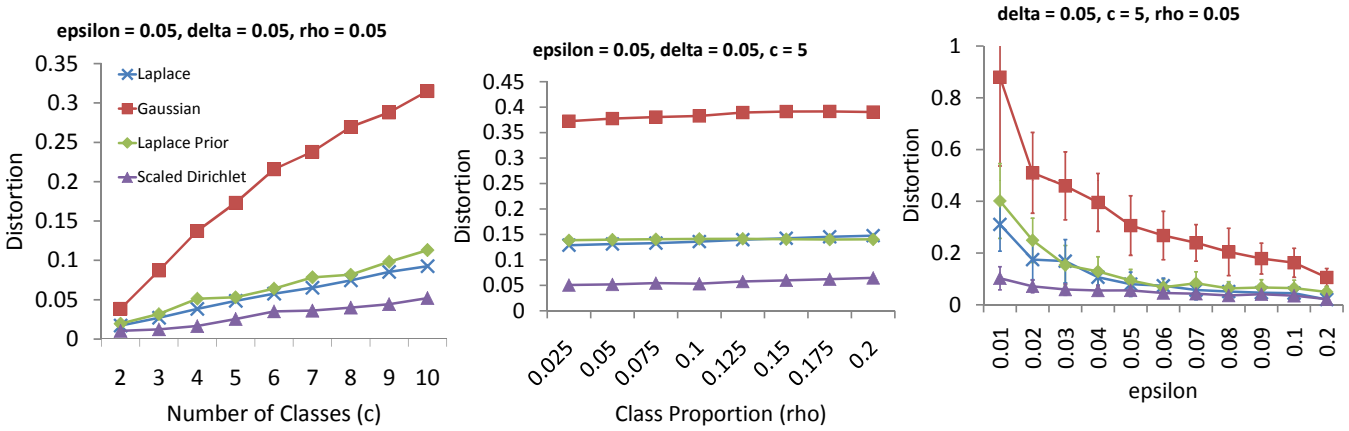

Figure 1: Comparing distortion of privacy mechanisms under varying settings of parameters. The first plot is for increasing number of classes, second plot for changing class proportions on five classes, third plot for increasing ϵ on five classes.

Varying number of classes.

The first plot in Figure 1 shows distortion in the output of different privacy mechanisms with increasing number of classes (c). We observe that as the number of classes increases, our method increasingly scores over all existing approaches. The key reason is that all other methods distort by adding independent noise to each class count. In contrast, our method preserves class ratios by using the same multiplicative parameter to increase variance when required.

Varying class proportions.

Next in the middle plot of Figure 1 we fix the number of classes to 5 and vary the class proportions ($\rho_1, \rho_1, \rho_1, \rho_1, 1-4\rho_1$) by varying ρ_1 from 0.025 (skewed proportions) to 0.2 (uniform proportions). We observe that the distortion achieved by our mechanism is the lowest under all settings.

Varying ϵ.

In the last plot in Figure 1 we vary ϵ from 0.01 to 0.2. As expected, the output is distorted more when the privacy requirements is more stringent (ϵ small). The Laplace mechanism is quite competitive with ours (Scaled Dirichlet) when ϵ is large but for small ϵ, our method provides much smaller distortion. Also, from the error bars we note that the variance of our method is the lowest.

We thus conclude that our privacy model has smaller distortion and smaller variance compared to other models, and is particularly useful when the number of classes is large.

4.2 Class ratio estimation models

In this section we compare various class ratio estimation models on set-labeled data without any perturbation of the class proportions.

We evaluate our method, which we call **MMD**, presented in Section 2.3 using a RBF kernel chosen through cross-validation. We vary the bandwidth of the kernels in the range 2^{-5} to 2^5. The only existing method that we are aware of that can be trained with set-labeled supervision and work with universal kernels is [28]. As discussed in Section 2, we can estimate class ratios using this method by aggregating per-instance predictions from the SVM classifier. We call this the **pSVM** method. The classifier we used was $\propto$SVM[7] [28] which trains a SVM like model from sets of instances and their class proportions. The parameters for this classifier are chosen via cross-validation. We vary the bandwidth of the kernels in the range 2^{-5} to 2^5. The other parameters of varied in the range as per [28].

Dataset	Number Features	Number Instances	$\lvert\mathcal{Y}\rvert$ (c)	n_i	n_u
Census	14	48,842	2	600	600
Youtube	1000	6,431,471	2	600	600
Higgs	28	11,000,000	2	600	600
Mnist	780	60,000	10	3000	3000
Acoustic	50	78,823	3	900	900
Twitter	3600	6,940	3	900	900

Table 1: Summary of Datasets

Datasets.

The method **pSVM** only works for binary classes. So, for the comparisons in this section we restrict to datasets with two classes. Table 1 summarizes the datasets we used.

Census: Census dataset consists of records from 1994 census database with features that include age, workclass, occupation, relationship, race, sex etc. The target label is income which is 1 if the person earns more than 50000 in a year and 0 otherwise. This dataset is available from the UCI repository[8].

Higgs: In a particle accelerator, not all collisions are likely to produce interesting particles. This dataset captures features of several collisions or processes. The label indicates whether a given collision or process is going to produce an interesting particle or not. 21 out of the 28 features are properties measured by the particle detectors in the accelerator and the remaining 7 are abstract features hand-designed by physicists. This dataset is available from the UCI repository.

Youtube: A dataset based on YouTube comments that we created based on this[9] collection. The goal in the YouTube dataset is to estimate the fraction of comments that are

[7]Code taken from https://github.com/felixyu/pSVM
[8]https://archive.ics.uci.edu/ml/datasets/Census+Income
[9]http://mlg.ucd.ie/yt

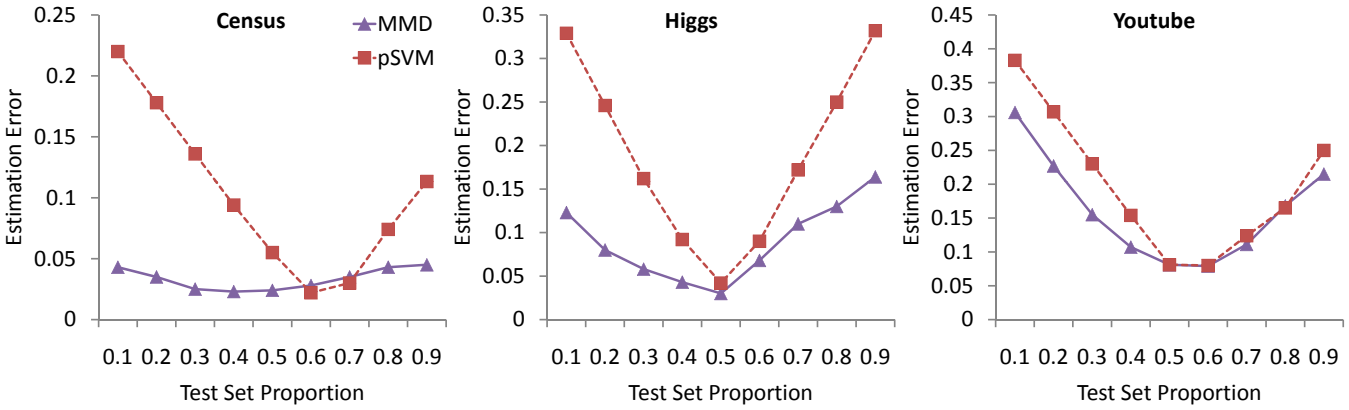

Figure 2: Comparison of models for class ratio estimation. Y-axis is the L1 difference between the true and estimated class ratios, and X-axis is the fraction of instances in class 1 in the test set. Each plot shows results for a different dataset. In all three plots, the method legends is the same.

spams on a YouTube video. The dataset was crawled by tracking 6407 popular YouTube videos over 77 days and comprises of 6,431,471 comments labeled spam or not. The feature set is a normalized TF-IDF vector over 1000 words + a comment length feature.

If the number of classes is c, we create c training sets in the following way: we initially choose a value ρ which is usually 0.05 or 0.1. Then, we create c sets with the following class proportions $[\rho, \rho, \ldots, 1-(c-1)\rho], [\rho, \rho, \ldots, 1-(c-1)\rho, \rho], [\rho, \ldots, 1-(c-1)\rho, \rho, \rho], \ldots, [1-(c-1)\rho, \rho, \ldots, \rho]$. The training sets are all class conditionally sampled. In addition, we add c more sets with the same class proportions as above to support cross validation for kernel selection.

Since, we are dealing with only binary datasets in this experiment, we create 4 sets in training, 2 of them have proportions $(0.1, 0.9)$ and the other 2 had proportions $0.9, 0.1$. This is similar to the process described above with $\rho = 0.1$. For testing, we created various test sets whose positive class proportions varied from 0.1 to 0.9.

For each method we measure estimation error as the L1 distance between the correct class proportions in the test set and the proportions estimated by the method. In Figure 2 we show the errors of the baseline **pSVM** and our **MMD** method on test sizes with increasing fraction of instances in the first class. We observe that **MMD** estimates have much lower error than **pSVM** especially for extreme class skews. The **pSVM** method assumes that the training and test distributions of $P(y)$ remain unchanged, and therefore this method returns accurate class ratios only in the range of ratios where this assumption holds. Our method allows the $P(y)$ distribution to shift and therefore provides lower error on a wider range.

We conclude from this section with set-labeled training data, our proposed **MMD** method provides the lowest error among existing options for a large range of test set ratios.

4.3 Private Class Ratio Estimation

In this section, we evaluate if our estimates continue to remain accurate when the label proportions are distorted to protect privacy. Our experiments in Section 4.1 showed that the distortion is within tolerable limits for binary datasets but is large for non-binary datasets. Therefore, in this section we report experiments on real-world non-binary datasets. The three we considered are described below and summarized in Table 1.

Mnist: This is a handwritten digit recognition dataset. The target labels are the digits 0 to 9 and inputs are fixed size input containing the handwritten image of a digit.

Acoustic: Acoustic is a three class dataset about classifying military vehicles from microphone recordings. This dataset as well as Mnist are available from LibSVM multiclass dataset repository[10].

Twitter: This twitter dataset was created for the task of classifying each tweet into one of three sentiment classes: positive, negative and neutral. We use the dataset and feature extraction mechanism described in [26]. The authors of [26] have made code and data available online[11].

In Figure 3 we show the **MMD** estimation error when trained with data distorted by three privacy mechanisms: ours (scaled dirichlet), Laplace, and Laplace Prior as defined in Section 4.1. We drop the Gaussian method since it provides no advantage over the Laplace method as per our experiments in Section 4.1. Instead we add as a reference the errors when **MMD** is trained with undistorted sets. Note this method does not guarantee any privacy and is just included to get a lower bound. The test set for these experiments where created with varying class proportions using the parameter ρ_1 as described in Section 4.1. The X-axis in Figure 3 indicates these class proportions.

We can make the following observations from these figures.

1. As expected, our estimation model is sensitive to the distortion of training data as revealed by the comparison with the "No distortion" setting. Thus it is important to devise mechanisms that reduce distortion. We showed in Section 4.1 that our method achieves the lowest distortion. These experiments show that the reduced distortion translates to reduced error of our learning algorithm.

2. When the number of classes is large (as in Mnist), existing methods (Laplace) introduce such large distortion that the errors of the learning model become unacceptably large.

[10] https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/multiclass.html

[11] https://github.com/duytinvo/ijcai2015

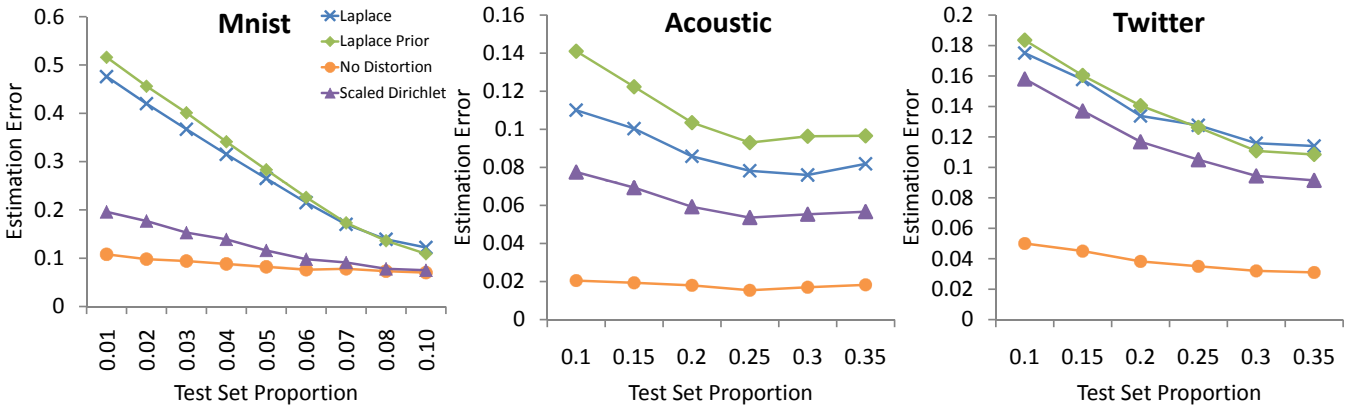

Figure 3: Our MMD estimation model trained with privacy protected data under three different mechanisms. For reference we also include error with correct set proportions (no privacy). The first plot is for Mnist, second for Acoustic and third for twitter. X-axis is increasing test set proportions.

3. The effect of bad training data is more pronounced when the test set has skewed proportions. This trend is very clear for the Mnist dataset where we see a sharp increase in error of Laplace distortion on highly skewed testset.

5. RELATED WORK

In this paper we presented a learning model for estimating proportions of labels in a set of instances using set-labeled training data with privacy constraints on the true label.

Our use of set-labeled data for training is related to the models proposed in [22, 28, 23, 25, 21] for training classification models with the same kind of supervision. However, the way we wrap an MMD objective around set-labeled data allows effective learning from a few, large sets. In contrast, the classification models of [28], while being kernel-based like us, prefer many small sets. Since small sets is anti-thetical to privacy, our method is particularly suitable for learning under privacy constraints.

The use of MMD for estimating class proportions has been explored before in [13, 29]. But these models require instance-labeled data during training which does not work under privacy constraints. One option would have been to train the model with private data while enforcing privacy during the model creation phase as used in [24, 4, 14, 16, 1] and several others [15]. We did not consider this option because we are targeting scenarios where the training data is aggregated from several private organizations and the model creation happens outside the trust boundary.

Another class of methods attempt to first create DP joint distribution of the data [6, 20] and then sample instances from the distribution for down-stream tasks like model-creation. These summaries are for general-purpose analysis, and not a specific classification task, and are expected to be less accurate. Ours and others like [21, 30] of creating summaries is geared for the prediction task at hand while allowing data from multiple agencies to be aggregated. We have shown that our mechanism for publishing label proportions provides much higher accuracy for the same (ϵ, δ) guarantees than existing approaches based adding Gaussian noise to each component as suggested in [8] and sanitizing the output using either the recently proposed constrained least square approach [18] or the Dirichlet samples [20].

A different category of approach attempt to publish combinatorial summaries of data, for instance [5] proposes to create trees to publish set-valued data for differential privacy. Even though the method is designed for publishing sets of items, their raw data is a large set of records each of which is small set of items. Consequently their definition of differential privacy (DP) is to be insensitive to the removal of a record. In contrast, our raw data is a single set of labels and our notion of DP is to be insensitive to the change of any single element of the set.

6. CONCLUSION

In this paper we designed a model for estimating class ratios and devised mechanisms for training it in scenarios where labels are provided on sets of instances and where labels are private. We theoretically analyzed our model and showed it to be consistent and accurate when the number of training sets is large. This is in contrast to existing methods that prefer many, small sets. Empirical evaluation on three real-world datasets show that our estimator provides lower error than existing methods, particularly when test class distributions are skewed.

We proposed a new mechanism for achieving differential privacy of labels that is more effective in preserving class ratios than existing mechanisms, particularly when the number of classes is large. We extend the learning model as well as its analysis for privacy-protected data. We show that the proposed learning and privacy mechanisms are well-suited for each other. In particular we show common conditions for achieving efficiency in both these phases. Empirical evaluation on several large real-datasets shows that the combination of our learning algorithm and privacy mechanism is able to provide significantly more accurate estimates than existing methods.

Our future work includes extending our model for the case of continuous y values (regression setting). Also, we would like to extend our mechanism to protect the privacy of the $\mathbf{x}$ part of an instance. Since our estimator is based on kernels, one idea is to use the technique of [24] for preserving privacy of the $\mathbf{x}$ and our current technique to protect the privacy of the ys.

7. REFERENCES

[1] R. Bassily, A. Smith, and A. Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *Foundations of Computer Science (FOCS), 2014 IEEE 55th Annual Symposium on*, pages 464–473. IEEE, 2014.

[2] A. Blum, C. Dwork, F. McSherry, and K. Nissim. Practical privacy: The sulq framework. In *PODS*, 2005.

[3] K. Chaudhuri and D. Hsu. Sample complexity bounds for differentially private learning. In *COLT*, 2011.

[4] K. Chaudhuri, C. Monteleoni, and A. D. Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12:1069–1109, 2011.

[5] R. Chen, N. Mohammed, B. C. M. Fung, B. C. Desai, and L. Xiong. Publishing set-valued data via differential privacy. *PVLDB*, 4(11), 2011.

[6] R. Chen, Q. Xiao, Y. Zhang, and J. Xu. Differentially private high-dimensional data publication via sampling-based inference. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 129–138. ACM, 2015.

[7] J. C. Duchi, M. I. Jordan, and M. J. Wainwright. Privacy aware learning. In *NIPS*. 2012.

[8] C. Dwork and A. Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014.

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

[10] A. Gretton, A. J. Smola, J. Huang, M. Schmittfull, K. M. Borgwardt, and B. Schölkopf. *Covariate shift and local learning by distribution matching*, pages 131–160. MIT Press, Cambridge, MA, USA, 2 2009.

[11] C. Guttmann, X. Sun, C. Rao, C. Queiroz, and B. I. Rubinstein. On the challenges of balancing privacy and utility of open health data. In *Joint Proceedings of the Workshop on AI Problems and Approaches for Intelligent Environments and Workshop on Semantic Cities*, pages 43–47. ACM, 2013.

[12] J. Hsu, M. Gaboardi, A. Haeberlen, S. Khanna, A. Narayan, B. C. Pierce, and A. Roth. Differential privacy: An economic method for choosing epsilon. In *IEEE 27th Computer Security Foundations Symposium, CSF*, 2014.

[13] 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.

[14] P. Jain and A. Thakurta. Differentially private learning with kernels. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pages 118–126, 2013.

[15] Z. Ji, Z. C. Lipton, and C. Elkan. Differential privacy and machine learning: a survey and review. *CoRR*, 2014.

[16] D. Kifer, A. Smith, and A. Thakurta. Private convex empirical risk minimization and high-dimensional regression. *JMLR*, 1:41, 2012.

[17] S. Le. *Learning via Hilbert space embedding of distributions*. PhD thesis, University of Sydney, School of Information Technologies, 2008.

[18] J. Lee, Y. Wang, and D. Kifer. Maximum likelihood postprocessing for differential privacy under consistency constraints. In *ACM SIGKDD*, 2015.

[19] Y. Li, K. Swersky, and R. S. Zemel. Generative moment matching networks. In *ICML*, 2015.

[20] A. Machanavajjhala, D. Kifer, J. M. Abowd, J. Gehrke, and L. Vilhuber. Privacy: Theory meets practice on the map. In *ICDE*, 2008.

[21] R. Nock, G. Patrini, and A. Friedman. Rademacher observations, private data, and boosting. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pages 948–956, 2015.

[22] G. Patrini, R. Nock, T. Caetano, and P. Rivera. (almost) no label no cry. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 190–198, 2014.

[23] N. Quadrianto, A. J. Smola, T. S. Caetano, and Q. V. Le. Estimating labels from label proportions. *Journal of Machine Learning Research*, 10:2349–2374, 2009.

[24] B. I. P. Rubinstein, P. L. Bartlett, L. Huang, and N. Taft. Learning in a large function space: Privacy-preserving mechanisms for svm learning. *Journal of Privacy and Confidentiality*, 4(1), 2012.

[25] S. Rüping. SVM classifier estimation from group probabilities. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10), June 21-24, 2010, Haifa, Israel*, pages 911–918, 2010.

[26] D.-T. Vo and Y. Zhang. Target-dependent twitter sentiment classification with rich automatic features. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, pages 1347–1353. AAAI Press, 2015.

[27] Y. Xin and T. Jaakkola. Controlling privacy in recommender systems. In *Advances in Neural Information Processing Systems*, pages 2618–2626, 2014.

[28] F. X. Yu, D. Liu, S. Kumar, T. Jebara, and S. Chang. $\propto$SVM for learning with label proportions. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, pages 504–512, 2013.

[29] K. Zhang, B. Schölkopf, K. Muandet, and Z. Wang. Domain adaptation under target and conditional shift. In *ICML*, 2013.

[30] S. Zhou, J. D. Lafferty, and L. A. Wasserman. Compressed and privacy-sensitive sparse regression. *IEEE Transactions on Information Theory*, 55(2):846–866, 2009.