Risk Minimization Under Sampling Bias Arising from Customer Interactions

Scott Rome* Michael Kreisel†

Abstract

In studying machine learning classifiers, researchers often assume that training and testing data are sampled at random from the same distribution. One way this assumption fails in practice is that training samples are biased, yielding training data drawn from a conditional distribution $p(\mathbf{x}, y \mid s = 1)$ rather than the true distribution $p(\mathbf{x}, y)$. In this paper, we consider the case of a call center where we are only able to collect the label y when a customer contacts us. This leads to a biased sampling model which depends on $\mathbf{x}$ only when $y = 1$. This sampling model is applicable to survey statistics and particularly data generated by voter surveys. By identifying a formal model for the sampling bias, we prove a generalization bound on the empirical risk of the optimal classifier f^s trained on the sampling distribution and characterize the tightness of this bound by the level of dependency between s and y and the empirical risk of the optimal classifier f^* on the full distribution.

Key Words: Sampling bias, generalization, statistical learning, machine learning, risk minimization, biased sample, data science

1. Introduction

In a typical setup for machine learning classification, one considers features $\mathbf{x}$ and labels y drawn from a distribution $D(\mathbf{x}, y)$. A model is then trained to estimate $p(y \mid \mathbf{x})$. However in many industrial settings, the data collection process is biased. We model this effect by adding a binary variable s which determines the selection of examples. Training data is drawn from $D(\mathbf{x}, y \mid s = 1)$ whereas testing data is drawn from $D(\mathbf{x}, y, s)$. A natural question is if insights derived from biased training data generalize to the larger population.

In [15], the authors discuss four ways that s can be related to $\mathbf{x}$ and y. The two non-trivial cases (listed as 2 and 3 in Section 2 of [15]) are when the selection variable s either depends only on the features $\mathbf{x}$ or only on the labels y. They examine the case when y is conditionally independent of s given $\mathbf{x}$ and give a procedure for reweighting examples which allows to train a classifier on biased data. Specifically, they describe a weighting of examples such that the estimated loss of a classifier on the weighted biased examples is equal to the estimated loss on unbiased data (Theorem 1 [15]).

In this paper, we prove an upper bound on the optimal classifier of the sample distribution that is similar in spirit to Theorem 1 of [15]. However, our assumptions on the relationship between $\mathbf{x}, y$, and s do not fit into any of their categories. We are motivated by the selection bias encountered when collecting training data via call centers and consider in particular the case of an internet service provider to ground the following discussion. Here $\mathbf{x}$ represents measurable features of a customer's service (like WiFi telemetry) and y represents whether they are experiencing a problem with that aspect of service (like slow WiFi speed). Although we can measure the state $\mathbf{x}$ continuously, we only observe y when a customer contacts us. Thus, our training

*Comcast Applied AI Research, 1800 Arch St, Philadelphia, PA 19103

†Comcast Applied AI Research, 1110 Vermont Ave NW, Washington DC, 20005

data to predict y given $\mathbf{x}$ is biased by the binary variable s which indicates whether the customer called.

1.1 Background

We assume that $\mathbf{x}$ is sampled from a probability space X, which in practice is a cartesian product with some continuous and categorical features. $D(\mathbf{x}, y, s)$ is a distribution on the product space $X \times \{0,1\} \times \{0,1\}$. The variable $s \in \{0,1\}$ indicates if a data point $(\mathbf{x}, y)$ has been sampled. We will use p to indicate the probability density function, and the notation $p(A, B) := p(A \cap B)$ to denote the joint probability distribution. For events A,B, and C, we say A and B are *conditionally independent given* C if $p(A, B|C) = p(A|C)p(B|C)$, which is equivalent to $p(A|B, C) = p(A|C)$.

Let ℓ denote a loss function $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ and $f : X \to \mathbb{R}$ a classifier. We define the ℓ-risk on D as

$$\mathbb{E}_D \ell(f(\mathbf{x}), y) := \sum_{\mathbf{x},y,s \in \mathcal{D}} \ell(f(\mathbf{x}, y))p(\mathbf{x}, y, s) = \sum_{\mathbf{x},y \in \mathcal{D}} \ell(f(\mathbf{x}, y))p(\mathbf{x}, y),$$

and note we have abused notation and written the integral as a sum. When we consider the risk over the sampled distribution $D(\mathbf{x}, y|s = 1)$, we will write

$$\mathbb{E}_D[\ell(f(\mathbf{x}), y)|s = 1] := \sum_{\mathbf{x},y \in \mathcal{D}} \ell(f(\mathbf{x}, y))p(\mathbf{x}, y|s = 1).$$

The risk minimizer of ℓ on $\mathcal{D}$ is defined to be $f^* := \text{argmin}_f \mathbb{E}_D \ell(f(\mathbf{x}), y)$.

1.1.1 Problem Statement

The following two conditions characterize the setting under study:

1. $\mathbf{x}$ and s are conditionally independent given $y = 0$.[1]
2. For any $\mathbf{x}, p(s = 1|\mathbf{x}, y = 1) \geq p(s = 1|\mathbf{x}, y = 0)$.

The conditions defined above do not fit into the classes described in [15]. Customer call in data provides an example where our assumptions reasonably may hold, but stronger assumptions as in [15] certainly do not.

For example, we argue that although $\mathbf{x}$ is not independent of s given $y, \mathbf{x}$ *is* conditionally independent of s given $y = 0$. When a customer contacts a call center, there are multiple scenarios that may have caused the interaction. For example, the customer may have a question about their bill or they may be calling to troubleshoot their WiFi. When the features $\mathbf{x}$ are specifically related to the service measured by y, if we know that the customer has no problem with their WiFi service ($y = 0$) then $\mathbf{x}$ has no bearing on whether the customer calls. On the other hand, given a set of features $\mathbf{x}$, a customer who calls is more likely to be experiencing problems than one who does not call. Thus $y = 1$ is not independent of s given $\mathbf{x}$.

Our second assumption says that customers with a problem ($y = 1$) are more likely to call than customers without a problem ($y = 0$) regardless of the value of $\mathbf{x}$. This assumption is intuitive and is not unique to the call center use case. For example, we might study surveys where y represents whether the participant intends to vote for a particular candidate, s represents whether the participant sends back their

[1] We note that this case does not necessitate that s and $\mathbf{x}$ are also conditionally independent given $y = 1$.

response, and $\mathbf{x}$ represents the demographics of the participant. At the discretion of the researcher, it may be reasonable to assume that a participant is more likely to respond to the survey if they intend to vote for the candidate. This amounts to making our second assumption.

Under these conditions, we will prove an upper bound on the generalization of the optimal classifier f^s trained on the call-in data, using characteristics of the sampling distribution and the optimal classifier f^* on the full distribution. Thus, even without adjusting the training data (as in [15]), we are able to train a classifier on the biased data which generalizes to the full distribution, and we characterize the penalty paid using the empirical risk for the biased data. We also provide further sufficient conditions under which f^s minimizes the risk on the full distribution.

1.2 Related Work

1.2.1 Sampling Bias and Label Noise

Theoretical approaches in modern machine learning to correct sampling bias stem from the study of cost-sensitive learning (e.g., [7, 16]). As discussed in the introduction, [15] originally identified sufficient conditions under which learning can occur under sampling bias using theoretical methods, which is a strong influence on this work. In particular, the sample weighting approaches identified in [15] have been extended to other fields such as the study of covariate shift [14], where they weight training set examples to mirror the testing set.

Our work builds on the literature by choosing a model for the sampling bias which is motivated by the call center use case. Given this model, we demonstrate that biased data is sufficient for training purposes in a practical setting. The techniques used to investigate the optimal classifier are reminiscent of the label noise literature, particularly [12, 8, 9, 12]. These studies typically compare the risk minimizer on the noise distribution to the risk minimizer on the true distribution. In our case, we are looking at the risk minimizer on the sampled distribution and compare it to the true distribution.

1.2.2 Connection to PU Learning

The underlying assumptions in our approach are similar to results found in PU Learning (for a review, see [2]). In PU Learning, the data contains only positive and unlabeled examples. In [6], they assume that the labeled data is "selected completely at random" (SCAR), where a subset of the positive class is labeled ($s = 1$) uniformly at random. They assume that s and $\mathbf{x}$ are conditionally independent given y and moreover that $p(s = 1|\mathbf{x}, y = 0) = 0$.

The SCAR assumptions are strictly stronger than our assumptions. Note that when $p(s = 1|\mathbf{x}, y = 0) = 0$ then trivially we have $p(s = 1|\mathbf{x}, y = 1) \geq p(s = 1|\mathbf{x}, y = 0)$. This illustrates a connection between the two problems and more generally the connection between PU Learning and supervised learning with sampling bias. PU learning can be employed in the call center use case if we ignore negative labels. For example, [3, 10] weight individual samples where [4, 5] identify special convex and non-convex loss functions for PU Learning. However, we prove that the addition of negatively labeled samples simplifies the training procedure in practice as opposed to PU Learning approaches. Without introducing bias through choosing a sampling scheme, Lemma 2.2 allows the practitioner not only to use a smaller sample but to

train via typical supervised learning and loss functions. Thus she also avoids crafting sampling weights which may add additional sources of error during training.

2. Risk Minimization Under Asymmetric Bias

2.1 Technical Lemmas

We first present two technical lemmas. We must make the following assumptions:

Assumption 2.1. *Assume* $\mathbf{x}$ *and* s *are conditionally independent given* $y = 0$.

Assumption 2.2. *For any* $\mathbf{x}$, $p(s = 1|\mathbf{x}, y = 1) \geq p(s = 1|\mathbf{x}, y = 0)$.

Lemma 2.1. *Under Assumptions 2.1 and 2.2, for all* $\mathbf{x}$

$$p(s = 1|\mathbf{x}) \geq p(s = 1|y = 0) \tag{1}$$

and

$$p(y = 1|\mathbf{x}, s = 1) \geq p(y = 1|\mathbf{x}). \tag{2}$$

Proof. We have

$$\begin{aligned} p(s = 1|\mathbf{x}) &= p(s = 1|\mathbf{x}, y = 0)p(y = 0|\mathbf{x}) + p(s = 1|\mathbf{x}, y = 1)p(y = 1|\mathbf{x}) \\ &= p(s = 1|y = 0)p(y = 0|\mathbf{x}) + p(s = 1|\mathbf{x}, y = 1)p(y = 1|\mathbf{x}) \\ &\geq p(s = 1|y = 0)p(y = 0|\mathbf{x}) + p(s = 1|y = 0)p(y = 1|\mathbf{x}) \\ &= p(s = 1|y = 0). \end{aligned}$$

We use Assumption 2.1 in the second equality and Assumption 2.2 in the inequality in line 3. Furthermore, since $p(s = 1|\mathbf{x}) \geq p(s = 1|y = 0)$, we can apply Bayes law to deduce

$$\begin{aligned} p(y = 1|s = 1, \mathbf{x}) &= 1 - p(y = 0|\mathbf{x}, s = 1) \\ &= 1 - \frac{p(s = 1|\mathbf{x}, y = 0)p(y = 0|\mathbf{x})}{p(s = 1|\mathbf{x})} \\ &= 1 - \frac{p(s = 1|y = 0)p(y = 0|\mathbf{x})}{p(s = 1|\mathbf{x})} \\ &\geq 1 - p(y = 0|\mathbf{x}) \\ &= p(y = 1|\mathbf{x}). \end{aligned}$$

□

Next we will prove that minimizing the loss on the sampled data set corresponds to minimizing the loss on the true data set up to a penalty depending on the underlying distribution.

Lemma 2.2. *Let* ℓ *be a nonnegative loss function and* h *a classification function. Denote by* α *the constant* $\frac{p(y=0)p(s=1)}{p(y=0,s=1)}$. *Under Assumptions 2.1 and 2.2 we have*

$$\alpha \mathbb{E}_{\mathbf{x},y\sim D}[\ell(h(\mathbf{x}), y)|s = 1] \geq \mathbb{E}_{\mathbf{x},y\sim D}\ell(h(\mathbf{x}), y).$$

Proof. When $y = 0$, we use Assumption 2.1 to compute

$$\begin{aligned} p(\mathbf{x}, y = 0|s = 1) &= p(\mathbf{x}|y = 0, s = 1)p(y = 0|s = 1) \quad (3) \\ &= p(\mathbf{x}|y = 0)p(y = 0|s = 1) \\ &= \frac{1}{\alpha}p(\mathbf{x}, y = 0). \quad (4) \end{aligned}$$

When $y = 1$, we apply inequalities (1) and (2) to yield

$$\begin{aligned} p(\mathbf{x}, y = 1|s = 1) &= p(y|\mathbf{x}, s = 1)p(\mathbf{x}|s = 1) \\ &\geq p(y = 1|\mathbf{x})p(\mathbf{x}|s = 1) \\ &= p(y = 1|\mathbf{x})p(\mathbf{x})\frac{p(s = 1|\mathbf{x})}{p(s = 1)} \\ &\geq p(y = 1|\mathbf{x})p(\mathbf{x})\frac{p(s = 1|y = 0)}{p(s = 1)} \\ &= \frac{1}{\alpha}p(\mathbf{x}, y = 1). \quad (5) \end{aligned}$$

By combining (4) and (5),

$$p(\mathbf{x}, y|s = 1) \geq \frac{1}{\alpha}p(\mathbf{x}, y).$$

Now we can bound the true loss

$$\begin{aligned} \alpha\mathbb{E}_{\mathbf{x},y\sim D}[\ell(h(\mathbf{x}), y)|s = 1] &= \alpha\sum_{\mathbf{x},y}\ell(h(\mathbf{x}), y)p(\mathbf{x}, y|s = 1) \\ &\geq \sum_{\mathbf{x},y}\ell(h(\mathbf{x}), y)p(\mathbf{x}, y) \\ &= \mathbb{E}_{\mathbf{x},y\sim D}\ell(h(\mathbf{x}), y). \end{aligned}$$

□

By averaging over $\mathbf{x}$ in (1) we can see that $\alpha \geq 1$. Thus we can only expect a larger discrepancy between $\mathbb{E}_{\mathbf{x},y\sim D}[\ell(h(\mathbf{x}), y)|s = 1]$ and $\mathbb{E}_{\mathbf{x},y\sim D}\ell(h(\mathbf{x}), y)$ when y and s are dependent.

Corollary 2.2.1. *Let ℓ be a nonnegative loss function. Suppose $D(\mathbf{x}, y|s = 1)$ is separable with respect to ℓ, so that there exists a classification function h with $\mathbb{E}_{\mathbf{x},y\sim D}[\ell(h(\mathbf{x}), y)|s = 1] = 0$. Then h also separates $D(\mathbf{x}, y)$ so that*

$$\mathbb{E}_{\mathbf{x},y\sim D}[\ell(h(\mathbf{x}), y)] = 0.$$

Proof. Follows immediately by plugging the optimal h into the bound from Lemma 2.2. □

Lemma 2.2 and Corollary 2.2.1 demonstrate that a model which is trained on the biased data will also generalize to the full distribution, albeit with a penalty of α on the loss. In the case when the distributions are separable, this penalty vanishes and the optimal classifier is the same for the biased and unbiased data. However even when the distribution is not separable, we would like some guarantee that the optimal classifiers on the biased and unbiased data are not too different. That is the subject of the next section.

2.2 Upper Bound on the Optimal Classifier of the Sampled Distribution

Let f^* denote the risk minimizer of ℓ for D, and f^s the risk minimizer of ℓ for $D(\mathbf{x}, y|s=1)$. Our goal is to show under what conditions f^s has risk bounded above by the risk of f^* on the full distribution D.

Theorem 2.3. *Let ℓ be a bounded, nonnegative loss function and f^*, f^s the risk minimizers for $D(\mathbf{x}, y)$ and $D(\mathbf{x}, y|s=1)$ respectively. Then under Assumptions 2.1 and 2.2, the following generalization bound holds:*

$$\begin{aligned}\mathbf{E}_{\mathbf{x},y}[\ell(f^s(\mathbf{x}), y)] &\leq \mathbf{E}_{\mathbf{x},y}[\ell(f^*(\mathbf{x}), y)] \\ &+\frac{p(y=1)}{p(s=1)}\mathbf{E}_{\mathbf{x}}\Big[\ell(f^*(\mathbf{x}), 1)\big(p(s=1|\mathbf{x}, y=1) - p(s=1|y=0)\big)\Big|y=1\Big].\end{aligned}$$

Proof. First, we will work to calculate an upper bound for $\mathbf{E}_{\mathbf{x},y}[\ell(f^*(\mathbf{x}), y))|s=1]$. In particular,

$$\begin{aligned}\mathbf{E}_{\mathbf{x},y}[\ell(f^*(\mathbf{x}), y))|s=1] &= \sum_{\mathbf{x},y}\ell(f^*(\mathbf{x}), y)p(\mathbf{x}, y|s=1) \\ &= \sum_{\mathbf{x}}\ell(f^*(\mathbf{x}), 0)p(\mathbf{x}, y=0|s=1) + \ell(f^*(\mathbf{x}), 1)p(\mathbf{x}, y=1|s=1).\end{aligned} \tag{6}$$

For the first term we use Assumption 2.1 to compute

$$p(\mathbf{x}, y=0|s=1) = \frac{1}{\alpha}p(\mathbf{x}, y=0) \tag{7}$$

as in equations (3) through (4). We next bound the second term of (6). We add and subtract $\sum_{\mathbf{x}}\frac{1}{\alpha}\ell(f^*(\mathbf{x}), 1)p(\mathbf{x}, y=1)$ to (6), yielding

$$\begin{aligned}&\mathbf{E}_{\mathbf{x},y}[\ell(f^*(\mathbf{x}), y))|s=1] \\ &= \frac{1}{\alpha}\sum_{x,y}\ell(f^*(\mathbf{x}), y)p(\mathbf{x}, y) \\ &+ \sum_{\mathbf{x}}\ell(f^*(\mathbf{x}), 1)\bigg(p(\mathbf{x}, y=1|s=1) - \frac{1}{\alpha}p(\mathbf{x}, y=1)\bigg).\end{aligned} \tag{8}$$

The first term is $\frac{1}{\alpha}\mathbf{E}_{\mathbf{x},y}[\ell(f^*(\mathbf{x}), y)]$. We define

$$\begin{aligned}\Gamma(\mathbf{x}) &:= p(\mathbf{x}, y=1|s=1) - \frac{1}{\alpha}p(\mathbf{x}, y=1) \\ &= \frac{p(\mathbf{x}, y=1)}{p(s=1)}\bigg(p(s=1|y=1, \mathbf{x}) - p(s=1|y=0)\bigg).\end{aligned} \tag{9}$$

Rearranging some terms, we have

$$\begin{aligned}&\sum_{\mathbf{x}}\ell(f^*(\mathbf{x}), y)\Gamma(\mathbf{x}) \\ &\quad= \frac{p(y=1)}{p(s=1)}\mathbf{E}_{\mathbf{x}}\Big[\ell(f^*(\mathbf{x}), 1)\big(p(s=1|\mathbf{x}, y=1) - p(s=1|y=0)\big)\Big|y=1\Big].\end{aligned} \tag{10}$$

Now, we may apply Lemma 2.2 to (8) and deduce the inequality

$$\begin{aligned}\mathbf{E}_{\mathbf{x},y}[\ell(f^s(x),1)] &\leq \alpha \mathbf{E}_{\mathbf{x},y}[\ell(f^s(x),y)|s=1] \\ &\leq \alpha \mathbf{E}_{\mathbf{x},y}[\ell(f^*(x),y)|s=1] \\ &\leq \mathbf{E}_{\mathbf{x},y}[\ell(f^*(x),y)] \\ &+\frac{p(y=1)}{p(s=1)}\mathbf{E}_{\mathbf{x}}\Big[\ell(f^*(\mathbf{x}),1)\big(p(s=1|\mathbf{x},y=1)-p(s=1|y=0)\big)\Big|y=1\Big]. \quad (11)\end{aligned}$$

□

This theorem provides an oracle bound in the case of biased data. The bound is not feasible to compute in practice, as it expects knowledge of both f^* and the full distribution. We see that the bound is controlled by the differences between $p(s=1|\mathbf{x},y=1)$ and $p(s=1|y=0)$ and we have equality if they are equal for all $\mathbf{x}$. This demonstrates that models trained on the biased sample will generalize to the full distribution if the bias is not too strong.

Now we present a corollary where we further simplify the bound (in a special case) to make it more interpretable and to deduce insight about the bias' impact.

Corollary 2.3.1. *Let ℓ be the $0-1$ loss. Under the assumptions of Theorem 2.3, we have the following bound:*

$$\begin{aligned}\mathbf{E}_{\mathbf{x},y}[\ell(f^s(\mathbf{x}),y)] &\leq \mathbf{E}_{\mathbf{x},y}[\ell(f^*(\mathbf{x}),y)] \\ &+\frac{p(y=1)}{p(s=1)}\Big(p(f^*(\mathbf{x})=0|y=1)\big(p(s=1|y=1)-p(s=1|y=0)\big) \\ &\quad + Cov\Big[f^*(\mathbf{x})=0, p(s=1|y=1,\mathbf{x})\Big|y=1\Big]\Big).\end{aligned}$$

Proof. Picking up from (11), we employ the identify on expectations,

$$E[XY] = E[X]E[Y] - Cov(X,Y)$$

to calculate

$$\begin{aligned}&\mathbf{E}_{\mathbf{x}}\Big[\ell(f^*(\mathbf{x}),1)\big(p(s=1|\mathbf{x},y=1)-p(s=1|y=0)\big)\Big|y=1\Big] \\ =&\mathbf{E}_{\mathbf{x}}\Big[\ell(f^*(\mathbf{x}),1)\Big|y=1\Big]\mathbf{E}_{\mathbf{x}}\Big[\big(p(s=1|\mathbf{x},y=1)-p(s=1|y=0)\big)\Big|y=1\Big] \\ &+Cov\Big[\ell(f^*(\mathbf{x}),1), p(s=1|\mathbf{x},y=1)\Big|y=1\Big] \\ :=&A+B. \quad (12)\end{aligned}$$

Notice by definition of the $0-1$ loss, $\ell(f^*(\mathbf{x}),1) := \mathbf{1}_{f^*(\mathbf{x})=0}$, and so we may simplify A by calculating the expectations:

$$\begin{aligned}A &= \mathbf{E}_{\mathbf{x}}\Big[\mathbf{1}_{f^*(\mathbf{x})=0}\Big|y=1\Big]\left(p(s=1|y=1)-p(s=1|y=0)\right) \\ &= p(f^*(\mathbf{x})=0|y=1)\left(p(s=1|y=1)-p(s=1|y=0)\right). \quad (13)\end{aligned}$$

Similarly, the covariance term is equivalent to

$$B = Cov\Big[f^*(\mathbf{x})=0, p(s=1|y=1,\mathbf{x})\Big|y=1\Big]. \quad (14)$$

□

In this case, we can see a clear connection between the underlying problem and the guarantee on training. From the corollary, we can see there are three conditions that must be met in order for training on the sampled data set to generalize well. First, the covariance between f^* and $p(s = 1|x, y = 1)$ should be small. This is interesting because f^* depends on $p(x, y)$ and not s; however, we expect s to also depend on $p(x, y)$. The degree in which these two relationships are similar can hinder the training of f^s. Next, either $p(f^*(x) = 0|y = 1) \approx 0$ or $p(s = 1|y = 1) \approx p(s = 1|y = 0)$ for generalization f^s. The first case can occur when the problem is nearly separable, but the latter case depends solely on the sampling process s.

2.2.1 Sufficient Conditions for Equality of f and f^*

Theorem 2.3 gives an upper bound on the risk, but under further assumptions we can show that f^s has equal risk to f^* on the full distribution D. We will identify a family of distributions with a parameter t where equality holds, and use this family to generate numerical examples in the next section. This will show that for the 0-1 loss, the optimal classifier on the biased data is also optimal for the full distribution. We will need the following assumptions:

Assumption 2.3. *For all pairs* $\mathbf{x}, \mathbf{x}', p(y = 1|\mathbf{x}) \leq p(y = 1|\mathbf{x}')$ *implies* $p(s = 1|\mathbf{x}) \leq p(s = 1|\mathbf{x}')$.

Assumption 2.4. *There exists an* $\mathbf{x}$ *such that* $p(s = 1|\mathbf{x}) = p(s = 1|y = 0)$.

Assumption 2.3 is a stronger version of Assumption 2.2. It says that when $\mathbf{x}$ is more indicative of a problem then it is more indicative of a call. Assumption 2.4 is a technical assumption needed for the proof.

Lemma 2.4. *Under Assumptions 2.1, 2.3, and 2.4 there exists a* $t \in \mathbb{R}$ *such that for all* $\mathbf{x}$, $p(y = 1|\mathbf{x}) \leq t$ *implies* $p(s = 1|\mathbf{x}) = p(s = 1|y = 0)$ *and* $p(y = 1|\mathbf{x}, s = 1) = p(y = 1|\mathbf{x})$.

Proof. The value $p(s = 1|y = 0)$ acts as the base sampling rate for our data. We denote it by the constant η. Define X to be the set

$$X := \{\mathbf{x}|p(s = 1|\mathbf{x}) = \eta\}$$

and define

$$t := \sup_{\mathbf{x} \in X} p(y = 1|\mathbf{x}).$$

By Assumption, 2.4 X is nonempty, so t is well defined.

Now, we claim that t has the required properties. First, assume that $\mathbf{x}$ satisfies $p(y = 1|\mathbf{x}) \leq t$. Then, by the definition of t, there exists $\mathbf{x}' \in X$ such that $p(y = 1|\mathbf{x}) \leq p(y = 1|\mathbf{x}')$. Applying our assumptions implies that $p(s = 1|\mathbf{x}) \leq p(s = 1|\mathbf{x}') = \eta$. By Lemma 2.1, $p(s = 1|\mathbf{x}) \geq \eta$, so we must actually have $p(s = 1|\mathbf{x}) = \eta$. We can compute

$$\begin{aligned} p(y = 1|\mathbf{x}, s = 1) &= 1 - \frac{\eta p(y = 0|\mathbf{x})}{p(s = 1|\mathbf{x})} \\ &= 1 - \frac{\eta p(y = 0|\mathbf{x})}{\eta} \\ &= p(y = 1|\mathbf{x}) \end{aligned}$$

which completes the proof. □

Theorem 2.5. *Let t be defined as in Lemma 2.4. Under the assumptions of Lemma 2.4, if $t \geq 1/2$ then*

$$\mathbb{E}_D \ell_{0,1}(f^s(\mathbf{x}), y) = \mathbb{E}_D \ell_{0,1}(f^*(\mathbf{x}), y).$$

Proof. We consider the set $S = \{\mathbf{x} : f^s(\mathbf{x}) \neq f^*(\mathbf{x})\}$. We will show $S = \emptyset$. First, due to Lemma 2.1, $p(y = 1|\mathbf{x}, s = 1) \geq p(y = 1|\mathbf{x})$. Thus, the only instances in S must be when $f^s(\mathbf{x}) = 1$ and $f^*(\mathbf{x}) = 0$. This implies any $\mathbf{x} \in S$ would satisfy $p(y = 1|\mathbf{x}, s = 1) \geq 1/2$ while $p(y = 1|\mathbf{x}) < 1/2$ due to the definition of the Bayes optimal classifier. However, $p(y = 1|\mathbf{x}) < 1/2 \leq t$, so $p(y = 1|\mathbf{x}, s = 1) = p(y = 1|\mathbf{x})$. Therefore, $f^*(\mathbf{x}) = f^s(\mathbf{x})$ for all such $\mathbf{x}$, implying $S = \emptyset$. □

We can extend this theorem by applying the theory of classification-calibrated losses ([1, 11, 13, 17]). Under suitable conditions these losses have minimizers that also exhibit low risk in the 0-1 loss. In particular, the hinge loss has an analytical solution which is equivalent to the Bayes classifier when $p(y = 1|\mathbf{x}) \neq \frac{1}{2}$. This implies the corollary:

Corollary 2.5.1. *Under the assumptions of Theorem 2.5, it follows that*

$$\mathbb{E}_D \ell_{Hinge}(f^s(\mathbf{x}), y) = \mathbb{E}_D \ell_{Hinge}(f^*(\mathbf{x}), y).$$

3. Numerical Experiments

In this section, we provide an experimental validation of the upper bound derived in Theorem 2.3 and investigate the relation of the optimal classifiers f^s and f^* defined on the sampled and full distribution respectively. We use the construction outlined in Theorem 2.5 to define a family of examples defined by a parameter $t \in (0, 1)$ where we expect the risk defined by f^* and f^s to be equal when $t \geq 1/2$.

We generate data satisfying Assumptions 2.1 and 2.2. Inspired by [9], we sample $\mathbf{x} \sim U(-1, 1)^{20} \subset \mathbb{R}^{20}$ and define $p(y = 1|\mathbf{x}) = \frac{1}{2}\left(\frac{\mathbf{x}\cdot w}{\max_{\mathbf{x}} \mathbf{x}\cdot w} + 1\right) \in [0, 1]$ where

$$w := [1/2^{10}, -1/4^{10}] = [1/2, \ldots, 1/2, -1/4, \ldots, -1/4] \in \mathbb{R}^{20}$$

and the notation ^{10} indicates repeating and $\cdot$ the inner product. We fix $t, \alpha > 0$ then define $p(s = 1|\mathbf{x}, y) = \alpha + \max(\{0, p(y = 1|\mathbf{x}) - t\})$ when $y = 1$ and α otherwise. Marginalizing over y gives $p(s = 1|\mathbf{x}) = \alpha + p(y = 1|\mathbf{x}) \max(\{0, p(y = 1|\mathbf{x}) - t\})$. By construction, the data is not separable and the Bayes risk is nonzero. In Figure 1, we have provided a plot of $p(y = 1|\mathbf{x})$ and $p(y = 1|\mathbf{x}, s = 1)$ to show the bias that arises during sampling.

For our experiments, we trained a classifier f on the biased distribution and evaluated on the full distribution. We also calculated the optimal loss values for both the biased and full distributions, along with the upper bound from Theorem 2.3. The results of the experiment may be viewed in Table 1. We let $\alpha = .2$ and trained on an equal sized sample ($n \approx 20000$) for the biased and true distribution. The held out test set for each experiment was sampled from the true distribution with $n = 20000$. As expected, the bound in Theorem 2.3 holds for all values of t tested. Additionally, as t gets larger we can see f, f^s, and f^* converging to the same loss values. As we expect from Theorem 2.5, we see for $t \geq 1/2$ that the loss values for the optimal classifiers trained on the biased and unbiased data are approximately equal.

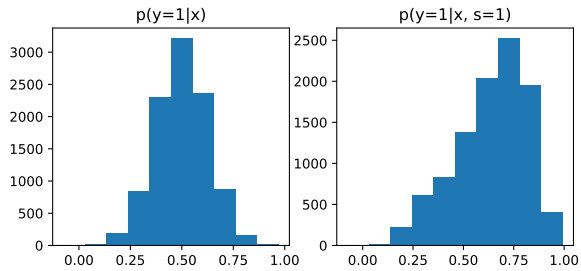

Figure 1: This plot compares the distributions of y from the sample to the full distribution. Here we sample $\mathbf{x} \sim U(-1, 1)^{20}$ and then plot a histogram corresponding to the calculated probability distribution. In this plot, $t = .3$ was used for all calculations. We can see that the sampled distribution of y differs significantly from the full distribution.

t	Hinge (Upper Bound)	Hinge (f^*)	Hinge (f^s)	Hinge (f)	Log Loss (Upper Bound)	Log Loss (f^*)	Log Loss (f^s)	Log Loss (f)
0.0	1.12	0.80	0.95	0.99	1.39	0.66	0.82	0.79
0.2	1.07	0.80	0.88	0.99	1.37	0.66	0.75	0.73
0.4	0.88	0.80	0.80	0.81	1.25	0.66	0.68	0.68
0.6	0.80	0.80	0.80	0.80	1.05	0.66	0.66	0.66
0.8	0.80	0.80	0.80	0.80	1.00	0.66	0.66	0.66

Table 1: Comparison of the loss values of f^*, f^s and f trained on sampled data for different t and the theoretical bound on the full distribution (denoted "Upper Bound" in the table). One can see that not only that f is close to the performance of f^*, but the upper bound from Theorem 2.3 holds as well.

4. Discussion

While most machine learning methods assume that the training and test sets are sampled from the same distribution, this assumption is often violated in practice. Motivated by data collected from call centers, we identify sufficient conditions for learning from biased data. Our analysis extends to all the common loss functions and distributions used in classification and regression tasks. Somewhat surprisingly, our analysis shows that in this case no adjustments need to be made to the biased data for models to generalize; however, depending on the underlying bias, one may not achieve the minimal risk on the full distribution. This is in contrast to Theorem 1 in [15] where weights must be obtained for each example, likely by training a second model. Our experiments validate our findings and demonstrate that learning is possible even when the training and test distributions are far from equal.

However we do pay a penalty for having biased data, as shown in the formula for the upper bound. The terms have intuitive interpretations as a measurement of the dependence between the events $y = 0$ and $s = 1$ and the performance of the optimal classifier f^* on the full distribution. Thus, as expected, it is never an advantage to train with biased data. In Theorem 2.3 the best case occurs when the events $y = 0$ and $s = 1$ are independent, which coincides with the standard conditions for machine learning.

4.1 Application to Surveys

Our approach can be applicable to other settings. For example, consider models to predict voter turnout based off of text message surveys sent to party members obtained through voter registration data. Let y indicate whether they plan to vote and $\mathbf{x}$ represent geographic and demographic factors. In this case, s is a measure of engagement separate from $\mathbf{x}$ (i.e. they responded to the text message survey). In this setting, $p(y = 1|\mathbf{x})$ can be interpreted as the citizen's enthusiasm about voting. The goal is to accurately identify $p(y = 1|\mathbf{x})$, but your only measurement of y is through a response $s = 1$. It is reasonable to assume that if they do not plan to vote ($y = 0$), then their demographics have no effect on their likelihood to respond $s = 1$, hence $\mathbf{x}$ and s are conditionally independent given $y = 0$. Moreover, it is reasonable to believe that voters are more likely to respond than non-voters. Thus, the conditions of Theorem 2.3 hold in this case.

References

[1] Bartlett, P. L., Jordan, M. I., and McAuliffe, J. D. Convexity, classification, and risk bounds. *Journal of the American Statistical Association 101*, 473 (2006), 138–156.

[2] Bekker, J., and Davis, J. Learning from positive and unlabeled data: A survey. *arXiv* (2018), https://arxiv.org/pdf/1811.04820v1.pdf.

[3] Bekker, J., and Davis, J. Learning from positive and unlabeled data under the selected at random assumption. In *Proceedings of the Second International Workshop on Learning with Imbalanced Domains: Theory and Applications* (ECML-PKDD, Dublin, Ireland, 10 Sep 2018), L. Torgo, S. Matwin, N. Japkowicz, B. Krawczyk, N. Moniz, and P. Branco, Eds., vol. 94 of *Proceedings of Machine Learning Research*, PMLR, pp. 8–22.

[4] du Plessis, M., Niu, G., and Sugiyama, M. Analysis of learning from positive and unlabeled data. In *NIPS* (2014).

[5] du Plessis, M., Niu, G., and Sugiyama, M. Convex formulation for learning from positive and unlabeled data. In *IMCL* (2015).

[6] Elkan, and Noto. Learning classifiers from only positive and unlabeled data. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 10*, 1145/1401890 (2008), 213–220.

[7] Elkan, C. The foundations of cost-sensitive learning. In *International joint conference on artificial intelligence* (2001), vol. 17, Lawrence Erlbaum Associates Ltd, pp. 973–978.

[8] Ghosh, A., Kumar, H., and Sastry, P. Robust loss functions under label noise for deep neural networks. In *AAAI* (2017), pp. 1919–1925.

[9] Ghosh, A., Manwani, N., and Sastry, P. Making risk minimization tolerant to label noise. *Neurocomputing 160* (2015), 93–107.

[10] Lee, W., and Liu, B. Learning with positive and unlabeled examples using weighted logistc regression. In *Procedings of the Twentieth International Conference on Machine Learning* (2003), pp. 448–455.

[11] Lugosi, G., Vayatis, N., et al. On the bayes-risk consistency of regularized boosting methods. *The Annals of statistics 32*, 1 (2004), 30–55.

[12] Manwani, N., and Sastry, P. Noise tolerance under risk minimization. *IEEE transactions on cybernetics 43*, 3 (2013), 1146–1151.

[13] Steinwart, I. Consistency of support vector machines and other regularized kernel classifiers. *IEEE Transactions on Information Theory 51*, 1 (2005), 128–142.

[14] Sugiyama, M., Suzuki, T., Nakajima, S., Kashima, H., von Bünau, P., and Kawanabe, M. Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics 60*, 4 (2008), 699–746.

[15] Zadrozny, B. Learning and evaluating classifiers under sample selection bias. *Proceedings of the International Conference on Machine Learning* (2004), 903–910.

[16] Zadrozny, B., Langford, J., and Abe, N. Cost-sensitive learning by cost-proportionate example weighting. In *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on* (2003), IEEE, pp. 435–442.

[17] Zhang, T. Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics* (2004), 56–85.