Journal of Machine Learning Research 18 (2018) 1-33 Submitted 5/2015; Revised 1/2017; Published 4/2018

# Cost-Sensitive Learning with Noisy Labels

**Nagarajan Natarajan** NAGARAJN@MICROSOFT.COM*
*Microsoft Research,*
*Bangalore 560001, INDIA*

**Inderjit S. Dhillon** INDERJIT@CS.UTEXAS.EDU
*Dept. of Computer Science*
*University of Texas at Austin*
*Austin, TX 78701*

**Pradeep Ravikumar** PRADEEPR@CS.CMU.EDU
*Machine Learning Dept.*
*Carnegie Mellon University*
*Pittsburgh, PA 15213*

**Ambuj Tewari** TEWARIA@UMICH.EDU
*Dept. of Statistics, and*
*Dept. of Electrical Engineering and Computer Science*
*University of Michigan*
*Ann Arbor, MI 48109*

**Editor:** Guy Lebanon

## Abstract

We study binary classification in the presence of *class-conditional* random noise, where the learner gets to see labels that are flipped independently with some probability, and where the flip probability depends on the class. Our goal is to devise learning algorithms that are efficient and statistically consistent with respect to commonly used utility measures. In particular, we look at a family of measures motivated by their application in domains where cost-sensitive learning is necessary (for example, when there is class imbalance). In contrast to most of the existing literature on consistent classification that are limited to the classical 0-1 loss, our analysis includes more general utility measures such as the AM measure (arithmetic mean of True Positive Rate and True Negative Rate). For this problem of cost-sensitive learning under class-conditional random noise, we develop two approaches that are based on suitably modifying surrogate losses. First, we provide a simple unbiased estimator of any loss, and obtain performance bounds for empirical utility maximization in the presence of i.i.d. data with noisy labels. If the loss function satisfies a simple symmetry condition, we show that using unbiased estimator leads to an efficient algorithm for empirical maximization. Second, by leveraging a reduction of risk minimization under noisy labels to classification with weighted 0-1 loss, we suggest the use of a simple weighted surrogate loss, for which we are able to obtain strong utility bounds. This approach implies that methods already used in practice, such as biased SVM and weighted logistic regression, are provably noise-tolerant. For two practically important measures in our family, we show

∗. The work in this manuscript was done when the author was a graduate student at the University of Texas at Austin.

that the proposed methods are competitive with respect to recently proposed methods for dealing with label noise in several benchmark data sets.

**Keywords:** class-conditional label noise, statistical consistency, cost-sensitive learning

## 1. Introduction

Learning from noisy training data is a problem of theoretical as well as practical interest in machine learning. In many applications such as learning to classify images, it is often the case that the labels are noisy. Even human labelers are susceptible to errors in labeling; for instance, certain image categories may be hard to discern. Designing learning algorithms that help maximize a desired performance measure in such noisy settings, and understanding their statistical consistency properties are the objectives of our current work.

One of the earliest known attempts at learning in the presence of label noise was by Bylander (1994) that concerned learnability of linear threshold functions (LTFs) in the Probably Approximately Correct (PAC) model. In particular, he showed that if the noise rate is uniform and if there is a sufficient margin under the clean distribution, then it is possible to PAC-learn LTFs. He also extended the result to a more realistic noise model called monotonic noise (Bylander, 1998), where the noise rate is allowed to vary per example, but is assumed to be a monotonic function of the distance of the example from the true hyperplane. Blum and Mitchell (1998), and later Cohen (1997) improved the PAC-learnability results of Bylander (1994) showing that linear threshold functions are efficiently learnable without the margin requirement in the uniform label noise model. A Bayesian approach to the problem of noisy labels is taken by Graepel and Herbrich (2000) and Lawrence and Schölkopf (2001). Cesa-Bianchi et al. (2011) focus on online learning algorithms where only unbiased estimates of the gradient of the loss are needed to provide guarantees for learning with noisy data. However, they consider a much harder noise model where instances *as well as* labels are noisy. Because of the harder noise model, they necessarily require multiple noisy copies per clean example and the unbiased estimation schemes also become fairly complicated, particularly for non-smooth classification losses such as the hinge loss.

In order to more clearly understand the impact of label noise, it is useful to consider a more natural and simpler formalism for label noise, where a random noise process corrupts the labels (Biggio et al., 2011), which otherwise arise from some "clean" distribution. There has been a long line of work in the theoretical machine learning community on such formalisms. Soon after the introduction of the noise-free PAC model, Angluin and Laird (1988) proposed the *random classification noise* (RCN) model where each label is flipped independently with some probability $\rho \in [0, 1/2)$. It is known (Aslam and Decatur, 1996; Cesa-Bianchi et al., 1999) that finiteness of the VC dimension characterizes learnability in the RCN model. Similarly, in the online mistake bound model, the parameter that characterizes learnability without noise — the Littlestone dimension — continues to characterize learnability even in the presence of random label noise (Ben-David et al., 2009). These results are for the so-called 0-1 loss: if the true label is $y \in \{-1, +1\}$ and the prediction is $t$, the 0-1 loss defined as $\ell_{0\text{-}1}(y, t) = 1_{\{yt \leq 0\}}$ (where $1_{\{P\}}$ denotes the indicator function that takes value 1 if the predicate $P$ is true or 0 otherwise), is a non-convex function of the prediction $t$. On the other hand, learning with convex losses has been addressed only under limiting assumptions like separability or uniform noise rates (Manwani and Sastry,

2013). A great deal of practical work has also been done on the problem; see, for instance, the survey article by Nettleton et al. (2010).

In this paper, we consider the *class-conditional* random label noise (abbreviated CCN) setting. Here, the data consists of iid samples drawn from a noisy version $D_\rho$ of an underlying "clean" distribution $D$, and where the noise rates depend on the class label. To the best of our knowledge, general results in this setting have not been obtained before. We note that developing guarantees in the presence of CCN label noise also has implications in varied partially-supervised settings such as learning from only positive and unlabeled data (Elkan and Noto, 2008), which can be cast under this setting. For the theoretical results presented in this work, we assume that the true noise rates (that characterize $D_\rho$) are known. In practice, one may use the domain knowledge to provide an estimate for noise rates (see Section 6.3), or use a plug-in estimator for noise rates such as the one prescribed by Scott (2015).

A key facet of the classification problem is the evaluation metric which captures discrepancy between the predicted label and the true label, and which we would want to minimize (or correspondingly, an evaluation utility measure which we would want to maximize). While classification accuracy is a popular utility measure, many other performance measures have also been considered in practice. One important family of measures constitutes cost-sensitive learning, and is motivated by applications and domains where misclassification cost could depend on the category of the example. For example, in disease diagnosis, false positives and false negatives often have very different associated impacts. Most, if not all, of the existing theoretical work on classification focuses on obtaining consistent learning algorithms for the 0-1 loss or its surrogates. In this paper, we consider a general class of utility measures that can be expressed as a linear combination of the entries of the "confusion matrix," namely, true positives, true negatives, false positives and false negatives.

Towards this problem of learning classifiers with respect to general utility measures and class conditional label noise, we develop two methods for suitably modifying *any given surrogate loss function* $\ell$, and show that minimizing the sample average of the modified proxy loss function $\tilde{\ell}$ leads to provable utility bounds where the utility is calculated on the clean distribution.

In our first approach, the modified or proxy loss is an unbiased estimate of the loss function associated with the utility measure of interest. The idea of using unbiased estimators is well-known in stochastic optimization (Nemirovski et al., 2009). Nonetheless, we bring out some important aspects of using unbiased estimators of loss functions for empirical utility maximization under CCN. In particular, we give a simple symmetry condition on the loss (enjoyed, for instance, by the Huber, logistic, and squared losses) to ensure that the proxy loss is also convex. Hinge loss does not satisfy the symmetry condition, and thus leads to a non-convex problem. We nonetheless provide a convex surrogate, leveraging the fact that the non-convex hinge problem is "close" to a convex problem (Theorem 12). This is strikingly different from the online learning setting (examined in Section 4) that requires only the expected loss to be convex.

Our second approach is based on the fundamental observation that the minimizer of the risk (i.e. probability of misclassification) under the noisy distribution differs from that of the clean distribution *only* in where it thresholds $\eta(x) = P(Y = 1|x)$ to decide the label. In order to correct for the threshold, we then propose a simple weighted loss function, where

the weights are label-dependent, as the proxy loss function. Our analysis builds on the notion of consistency of weighted loss functions studied by Scott (2012). This approach leads to a remarkable result that appropriately weighted losses like biased SVMs studied by Liu et al. (2003) are robust to CCN.

The key contributions of the paper are summarized below:

1. We develop methods for learning, that are provably consistent, (a) in the presence of asymmetric label noise, and (b) with respect to general cost-sensitive utility measures beyond the classical 0-1 loss.

2. To the best of our knowledge, we are the first to provide guarantees for cost-sensitive learning under random label noise in the general setting of convex surrogates, without any assumptions on the true distribution.

3. As one consequence of our results, we resolve an elusive theoretical gap in the understanding of practical methods like biased SVM and weighted logistic regression: as we show, they are provably noise-tolerant (Theorem 18). We obtain the result as a consequence of being able to linearly relate the risk w.r.t. a weighted 0-1 loss under the noisy distribution to that w.r.t. the 0-1 loss under the clean distribution (Theorem 16).

4. Our proxy losses are easy to compute: the proposed approaches yield efficient algorithms.

5. Experiments on benchmark data sets show that the methods are robust even at high noise rates, for maximizing different performance measures from our family.

In a preliminary version of the paper (Natarajan et al., 2013), we provided guarantees for risk minimization (using the 0-1 loss) in the presence of class-conditional label noise. In this paper, we provide a more general and detailed treatment of the theory, by characterizing the optimal classifiers under more general performance measures used in practice (in Sections 4 and 5). We extend our earlier approach (Natarajan et al., 2013) to cost-sensitive learning (Section 3.2). Our results in this paper also serve to generalize some of the known consistency results for performance measures such as the AM measure, even in the noise-free setting.

We now expand on the organization of the paper. We begin by discussing some closely related work in Section 2. We introduce and set the problem up formally in Section 3. The class-conditional noise model is specified by two parameters $\rho_{+1}$ and $\rho_{-1}$ which correspond to the rates at which positive and negative labels are flipped (independently) respectively. To build the theory, we assume that the rates are known to the learner. We do not make any assumptions on the underlying distribution. We introduce the family of measures $\mathcal{U}$ that constitute cost-sensitive learning in Section 3. It is well-known that the classical 0-1 loss is optimized by thresholding $P(Y = 1|x)$ at $1/2$. Cost-sensitive measures are of particular interest to our work in this paper because their optimal decision function exhibits a simple form — thresholding the conditional probability $P(Y = 1|x)$ at a certain value (stated in Lemma 2). Study of consistent learning algorithms, for many performance measures other than classification accuracy, is limited even in the noise-free case. Menon et al. (2013)

showed consistency of certain empirical estimation algorithms for the AM measure (defined in Proposition 1). An important result that connects the excess risk of a decision function (in terms of the 0-1 loss), $R(f) - \min_f R(f)$, and its "utility deficit", $\max_f \mathcal{U}(f) - \mathcal{U}(f)$, is established in Lemma 3. As a consequence of this result, we are able to use the surrogates for 0-1 loss for empirical estimation, in order to maximize cost-sensitive measures.

We describe our first approach of using unbiased surrogate loss functions in Section 4. Here, the idea is to construct an unbiased estimator of a given loss function (a surrogate of the 0-1 loss). The unbiased estimator involves the noise rates $\rho_{+1}$ and $\rho_{-1}$. For optimizing a given utility measure $\mathcal{U}$, we propose an empirical risk minimization procedure based on the unbiased surrogate thus obtained. We establish utility deficit bounds for the resulting empirical estimator in Theorem 9. Here, we also look at the online learning setting, where examples arrive sequentially (with noisy labels), and obtain similar consistency guarantees. Our second approach is detailed in Section 5. The key observation is that the optimal decision function for utility $\mathcal{U}$ in our family with respect to the noisy distribution is simply given by thresholding $P(Y = 1|x)$ with respect to the *clean* distribution, at a certain value that depends only on the distribution and the measure $\mathcal{U}$ itself. This enables us to use a weighted surrogate of the 0-1 loss, where the weights depend on the measure $\mathcal{U}$ and noise rates $\rho_{+1}$ and $\rho_{-1}$. We provide rigorous guarantees for consistency of the resulting empirical estimator in Theorem 18.

We present detailed experimental results that support our theory in Section 6. We perform experiments on synthetic and benchmark data sets, on both the proposed approaches. We compare to state-of-the-art algorithms for learning with noisy data on different data sets and different noise settings. We use two performance measures in experiments: classification accuracy and the AM measure, as representatives of the family of measures considered in the paper.

## 2. Related Work

Stempfel and Ralaivola (2009) propose minimizing an unbiased proxy for the case of the hinge loss. However the hinge loss leads to a non-convex problem. Therefore, they propose heuristic minimization approaches for which no theoretical guarantees are provided. We address the issue in Section 4.1. As Adaboost is very sensitive to label noise, random label noise has also been considered in the context of boosting. Freund (2009) proposes a boosting algorithm based on a non-convex potential that is empirically seen to be robust against random label noise. Long and Servedio (2010) prove that any method based on a convex potential is inherently ill-suited to random label noise. Biggio et al. (2011) consider robust SVM formulation in the presence of random and adversarial label noise. However, they do not provide any theoretical justification. Practitioners have developed several noise tolerant versions of the perceptron algorithm, although many are heuristic and are not known to be provably robust. This includes the passive-aggressive family of algorithms (Crammer et al., 2006), confidence weighted learning (Dredze et al., 2008), AROW (Crammer et al., 2009) and the NHERD algorithm (Crammer and Lee, 2010). The survey article by Khardon and Wachman (2007) provides an overview of some of this literature. To the best of our knowledge, there are no known mistake-bounded perceptron algorithms under asymmetric label noise.

Manwani and Sastry (2013) consider whether empirical risk minimization of the loss itself on the noisy data is a good idea when the goal is to obtain small risk under the clean distribution. But the answer is affirmative only for 0-1 and squared losses. Therefore, if empirical risk minimization over noisy samples has to work, we necessarily have to change the loss used to calculate the empirical risk. More recently, Ghosh et al. (2014) prove that a loss function $\ell$ satisfying the symmetry condition $\ell(f(\mathbf{x}),1)+\ell(f(\mathbf{x}),-1)=C, \forall \mathbf{x}, \forall f$ for some constant $C$ are noise-tolerant, under the assumption that the classes are separable under the clean distribution (here, $\ell$ is said to be *noise-tolerant* if $E_{(X,Y)\sim D}\big[\ell_{0\text{-}1}(f^*_\ell(X),Y)\big] = E_{(X,Y)\sim D}\big[\ell_{0\text{-}1}(\tilde{f}^*_\ell(X),Y)\big]$, where $f^*_\ell$ and $\tilde{f}^*_\ell$ denote the minimizers of $\ell$-risk under clean and noisy distribution respectively). Furthermore, they show that by choosing a sufficiently large value of a parameter in the loss functions such as sigmoid loss, ramp loss and probit loss, the losses can be made tolerant to non-uniform label noise (i.e. noise rate is allowed to depend on the example) as well. Unfortunately, the aforementioned loss functions are all non-convex, and convex losses used in practice do not satisfy the sufficiency conditions. It remains an open question if the symmetry condition is indeed necessary for noise tolerance, at least under the separability assumption.

van Rooyen and Williamson (2015) extend the idea behind the method of unbiased estimators to more general learning settings beyond binary classification. For example, they consider semi-supervised learning, classification with more than two classes, and learning with partial labels (a partial label is a *set* of labels containing the true label).

Scott et al. (2013) also study the problem of learning classifiers under the class-conditional noise model. However, they approach the problem from a different set of assumptions — the noise rates are *not* known, and the true distribution satisfies a certain "mutual irreducibility" property. They model the observed noisy instances as arising from "contaminated" mixtures of positive and negative classes and show that the mixture proportions can be consistently estimated by *maximal denoising* of the noisy distributions. Blanchard and Scott (2014) establish similar results for the multi-class classification problem. Scott (2015) provides a consistent estimator with convergence rates for the cost parameter $\alpha$ in our weighted surrogate loss (Equation 2). In this paper, however, we select $\alpha$ by cross-validation and also examine the sensitivity of selecting $\alpha$ (in Section 6.3).

## 3. Preliminaries

Let $D$ be the underlying true distribution generating $(X,Y) \in \mathcal{X} \times \{\pm 1\}$ pairs from which $n$ iid samples $(X_1,Y_1),\ldots,(X_n,Y_n)$ are drawn. Let $\eta(X) = P(Y=1|X)$ under $D$.

### 3.1 Class-conditional Noise

After injecting random classification noise (independently for each $i$) into these samples, corrupted samples $(X_1,\tilde{Y}_1),\ldots,(X_n,\tilde{Y}_n)$ are obtained. The class-conditional random noise model (CCN, for short) is given by:

$$\begin{aligned} P(\tilde{Y}=-1|Y=+1) &= \rho_{+1}, \\ P(\tilde{Y}=+1|Y=-1) &= \rho_{-1}, \text{ and} \\ \rho_{+1}+\rho_{-1} &< 1. \end{aligned}$$

The corrupted samples are what the learning algorithm sees. We will assume that the noise rates $\rho_{+1}$ and $\rho_{-1}$ are known to the learner. Let the distribution of $(X, \tilde{Y})$ be $D_\rho$. Noisy labels are denoted by $\tilde{y}$. Let $\tilde{\eta}(X) = P(\tilde{Y} = 1|X)$ under $D_\rho$.

### 3.2 Cost-sensitive Classification

Let $f : \mathcal{X} \to \mathbb{R}$ denote a real-valued decision function. The goal in classification is to learn $f$ from a training sample, such that some cost or loss measure is minimized. The most common measure is the probability of misclassification, also called the **risk**, which is simply the expected 0-1 loss defined as

$$R(f) := R_D(f) := \mathbb{E}_{(X,Y)\sim D}\left[1_{\{\text{sign}(f(X))\neq Y\}}\right].$$

Minimizing the 0-1 loss on a training sample, over some class of decision functions, is often intractable. In practice, it is common to minimize a *surrogate* loss function that is chosen for its computational advantages such as convexity. Minimizing risk (or equivalently, maximizing the accuracy) of a classifier is however not always appropriate, and in fact practitioners have devised many alternative performance metrics to address specific needs of a target domain. Class imbalance is an important scenario where accuracy of classifier is not a good metric to optimize: a trivial classifier that assigns all the examples to the majority class will have a high accuracy. However, little is known about optimal classification or consistent algorithms for binary classification w.r.t. general performance measures, even when the observations are noise-free. An important family of performance measures that is preferred in many scenarios including heavy class imbalance and asymmetry in real-world costs associated with specific classes constitutes *cost-sensitive* learning. Cost-sensitive performance measures are given by a weighted combination of the four fundamental population quantities associated with the "confusion matrix" - true positives, false positives (also known as type-I error), false negatives (also known as type-II error) and true negatives as defined below:

$$TP(f; D) = E_{(X,Y)\sim D}[1_{\{f(X)=1,Y=1\}}], \quad TN(f; D) = E_{(X,Y)\sim D}[1_{\{f(X)=-1,Y=-1\}}]$$

$$FP(f; D) = E_{(X,Y)\sim D}[1_{\{f(X)=1,Y=-1\}}], \quad FN(f; D) = E_{(X,Y)\sim D}[1_{\{f(X)=-1,Y=1\}}].$$

We consider the family of measures $\mathcal{U}$ defined by:

$$\mathcal{U}(f; D) = a_{11}TP(f; D) + a_{10}FP(f; D) + a_{01}FN(f; D) + a_{00}TN(f; D), \tag{1}$$

given constants $a_0, a_{11}, a_{10}, a_{01}, a_{00}$ (that could depend on $D$). In the remainder of the paper, $\mathcal{U}$ refers to a measure in this family, unless specified otherwise. We will use the terms *performance* measure and *utility* measure interchangeably in this paper.

Next, we state two important, commonly-used measures in this family.

**Proposition 1** *1. The Accuracy measure $\mathcal{U}_{Acc}(f; D)$ belongs in the family (1) with $a_{11} = a_{00} = 1$ and $a_{10} = a_{01} = 0$.*

*2. The AM (Arithmetic Mean of TPR and TNR) measure defined as*

$$\mathcal{U}_{AM}(f; D) := \frac{1}{2}\big(TPR(f; D) + TNR(f; D)\big),$$

*where* $TPR(f;D) = P(f(X) = 1|Y = 1)$ *is the true positive rate and* $TNR(f;D) = P(f(X) = -1|Y = -1)$ *is the true negative rate, belongs in the family* (1) *with constants* $a_{10} = a_{01} = 0$, $a_{11} = \frac{1}{2(1-\pi)}$ *and* $a_{00} = \frac{1}{2\pi}$, *where* $\pi = P(Y = 1)$ *under* $D$.

#### 3.2.1 Cost-sensitive Classification without Label Noise

Before we present our approaches for cost-sensitive classification under class-conditional label noise, it will be useful to consider the setting without such label noise, and setup appropriate notation. Given a utility measure $\mathcal{U}$ and training data, our goal is to learn a decision function $f$ that maximizes $\mathcal{U}$ with respect to the clean distribution. The optimal decision function (called Bayes optimal) that maximizes $\mathcal{U}$ over all real-valued decision functions is denoted as $f^\star(x) := \arg\max_f \mathcal{U}(f;D)$. We denote by $\mathcal{U}^*$ the optimal utility value, i.e. $\mathcal{U}^* = \mathcal{U}(f^\star)$. It is not always possible to characterize the Bayes optimal of arbitrary performance measures. Cost-sensitive measures are particularly interesting because their Bayes optimal exhibits a simple form, and as a consequence, consistent algorithms are readily obtained in practice in the noise-free case. Bayes optimal classifier for the family (1) is characterized in the following Lemma. Recall that $\eta(x) = P(Y = 1|x)$ under $D$.

**Lemma 2** *The Bayes optimal of any measure* $\mathcal{U}$ *in family* (1) *is given by*

$$\arg\max_f \mathcal{U}(f;D) = \text{sign}(\eta(x) - \delta^*_D),$$

*where the threshold is defined as*

$$\delta^*_D = \frac{a_{00} - a_{10}}{a_{00} - a_{10} - a_{01} + a_{11}}.$$

The proof is simple and can be found elsewhere (Elkan, 2001). It is well-known that accuracy $\mathcal{U}_{Acc}(f;D)$ is maximized by $\text{sign}(\eta(x) - 1/2)$ which is also readily obtained by applying the above lemma. For the AM measure $\mathcal{U}_{AM}(f;D)$, one easily verifies that the threshold is $\pi = P(Y = 1)$.

For any general measure $\mathcal{U}$, we are interested in controlling the *deficit utility* which is $\mathcal{U}^* - \mathcal{U}(f;D)$. The following simple lemma relates the deficit utility for the family (1) to that of a certain weighted 0-1 risk.

**Lemma 3** *Define* $\alpha$*-weighted risk under distribution* $D$ *as:*

$$R_\alpha(f) := R_{\alpha,D}(f) := E_{(X,Y)\sim D}\Big[(1-\alpha)1_{\{Y=1\}}1_{\{f(X)\leq 0\}} + \alpha 1_{\{Y=-1\}}1_{\{f(X)>0\}}\Big].$$

*For any measure* $\mathcal{U}$ *in the family* (1)*:*

$$R_{\delta^*_D}(f) - R^*_{\delta^*_D} = \frac{1}{(a_{11} + a_{00}) - (a_{10} + a_{01})}(\mathcal{U}^* - \mathcal{U}(f;D)),$$

*where* $R^*_{\delta^*_D} = \min_f R_{\delta^*_D}(f)$.

**Proof** Let $c_1 = (a_{11}+a_{00})-(a_{10}-a_{01})$ and $c_2 = a_{00}-a_{10}$. From Lemma 2, we know $\delta_D^* = \frac{c_2}{c_1}$. Note that $1 > c_1 > 0$ for otherwise maximizing $\mathcal{U}$ would not make sense (See Remark 4), and therefore $0 \leq \delta_D^* \leq 1$. For any $f$, let $\theta$ denote the classifier $\theta(x) = \text{sign}(f(x))$. We can rewrite $\mathcal{U}(\theta)$ as $\mathcal{U}(\theta) = c_1[(1-\delta_D^*)TP + \delta_D^* TN] + \tilde{A}$, where $\tilde{A}$ is a constant. We have:

$$\begin{aligned}
R_{\delta_D^*}(\theta) &= E_{(X,Y)\sim D}\Big[\big((1-\delta_D^*)1_{\{Y=1\}} + \delta_D^* 1_{\{Y=0\}}\big).1_{\{\theta(X)\neq Y\}}\Big] \\
&= (1-\delta_D^*)P(Y=1,\theta(X)=-1) + \delta_D^* P(Y=-1,\theta(X)=1) \\
&= (1-\delta_D^*)FN + \delta_D^* FP \\
&= (1-\delta_D^*)(\pi - TP) + \delta_D^*(1-\pi-TN) \\
&= (1-\delta_D^*)\pi + \delta_D^*(1-\pi) - \big((1-\delta_D^*)TP + \delta_D^* TN\big) \\
&= (1-\delta_D^*)\pi + \delta_D^*(1-\pi) + \frac{\tilde{A}}{c_1} - \frac{1}{c_1}\mathcal{U}(\theta).
\end{aligned}$$

Observing that $(1-\delta_D^*)\pi + \delta_D^*(1-\pi) + \frac{\tilde{A}}{c_1}$ is a constant independent of $\theta$, the proof is complete. ■

**Remark 4** *Note that we can assume $(a_{11}+a_{00})-(a_{10}+a_{01}) > 0$, otherwise maximizing $\mathcal{U}$ would not make sense. If indeed, $(a_{11}+a_{00})-(a_{10}+a_{01}) < 0$, then Lemma 3 still holds but with $\mathcal{U}^*$ interpreted as $\mathcal{U}^* = \min_f \mathcal{U}(f;D)$.*

Of course, minimizing the $\alpha$-weighted risk on a training sample is not tractable. Scott (2012) extends the notion of the classification calibration defined by Bartlett et al. (2006) for the (unweighted) 0-1 loss. The following result of Scott (2012) tells us that by using a similarly weighted surrogate loss function $\ell_\alpha$, one can control the excess $\alpha$-weighted risk. Define $\ell_\alpha$-risk, $R_{\ell_\alpha,D}(f) = E_{(X,Y)\sim D}[\ell_\alpha(f(X),Y)]$, and $R^*_{\ell_\alpha} = \min_f R_{\ell_\alpha,D}(f)$.

**Lemma 5 ($\alpha$-classification calibration (Scott, 2012))** *Given a loss function $\ell(t,y)$, and $\alpha \in (0,1)$, define the $\alpha$-weighted loss:*

$$\ell_\alpha(t,y) = \big((1-\alpha)1_{\{y=1\}} + \alpha 1_{\{y=-1\}}\big)\ell(t,y)\ . \tag{2}$$

*$\ell_\alpha$ is $\alpha$-classification calibrated (or $\alpha$-CC) iff there exists a convex, non-decreasing and* invertible *transformation $\psi_{\ell,\alpha}$, with $\psi_{\ell,\alpha}(0) = 0$, such that*

$$\psi_{\ell,\alpha}(R_\alpha(f) - R^*_\alpha) \leq R_{\ell_\alpha,D}(f) - R^*_{\ell_\alpha}\ .$$

*In other words, consistency with respect to $\ell_\alpha$-risk implies consistency with respect to $\alpha$-weighted (0-1) risk for $\alpha$-CC losses. Also, for any $\ell$ that is classification-calibrated (Bartlett et al., 2006) (such as logistic, hinge and squared losses), the corresponding $\ell_\alpha$ is $\alpha$-CC.*

If we choose $\alpha = \delta_D^*$, Lemmas 3 and 5 together guarantee the consistency of using a weighted surrogate loss function in practice, when we obtain samples from the clean distribution $D$. However, if the labels are noisy, the outlined procedure is no longer consistent. One necessarily has to change the loss function $\ell$ or rather $\ell_\alpha$ to be able to tolerate the noise, as described in the next two sections.

**Remark 6** *Most commonly used loss functions such as hinge and logistic losses are (even) margin losses, i.e.* $\ell(t,y) = \phi(ty)$*, for some* $\phi : \mathbb{R} \to [0,\infty)$*. We could also consider an* uneven margin *loss function of the form:*

$$\ell(t,y) = 1_{\{y=1\}}\phi(t) + 1_{\{y=-1\}}\beta\phi(-\gamma t),$$

*for* $\beta, \gamma > 0$*. Scott (2012) showed that for convex* $\phi$*, the above defined uneven margin loss* $\ell$ *is classification-calibrated, and in turn, the corresponding* $\ell_\alpha$ *is* $\alpha$*-CC, when* $\beta = \frac{1}{\gamma}$*. Such uneven margin losses have been used in practice mostly as heuristics as pointed out in (Scott, 2012). Thus, in principle, we could use uneven margin losses, and all the results in this manuscript will hold just the same.*

**Notation.** We use letters with the 'tilde' accent to denote noisy versions of quantities or variables, e.g. $\tilde{\ell}$ is the loss function to be used on the noisy data, and $\tilde{y}$ denotes a noisy label. We use $\mathcal{F} : \mathcal{X} \to \mathbb{R}$ to denote a fixed class of real-valued decision functions. If $f$ is not quantified in a minimization, then it is implicit that the minimization is over all measurable functions. Instances are denoted by $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^d$. Though most of our results apply to a general function class $\mathcal{F}$, we instantiate $\mathcal{F}$ to be the set of hyperplanes of bounded $L_2$ norm, $\mathcal{W} = \{\mathbf{w} \in \mathbb{R}^d : \|\mathbf{w}\|_2 \leq W_2\}$ for certain specific results.

## 4. Approach of Unbiased Surrogates

The method of unbiased surrogates uses the noise rates to construct an unbiased estimator $\tilde{\ell}(t,\tilde{y})$ for the loss $\ell(t,y)$. The following key lemma tells us how to construct unbiased estimator of the loss from noisy labels.

**Lemma 7** *Let* $\ell(t,y)$ *be any bounded loss function. Then, if we define,*

$$\tilde{\ell}(t,y) := \frac{(1-\rho_{-y})\,\ell(t,y) - \rho_y\,\ell(t,-y)}{1-\rho_{+1}-\rho_{-1}}$$

*we have, for any* $t, y$*,* $\mathbb{E}_{\tilde{y}}\left[\tilde{\ell}(t,\tilde{y})\right] = \ell(t,y)$ *. In particular, for any given* $\alpha \in (0,1)$*,* $\mathbb{E}_{\tilde{y}}\left[\tilde{\ell}_\alpha(t,\tilde{y})\right] = \ell_\alpha(t,y)$*, where* $\ell_\alpha(t,y)$ *is defined as in* (2)*.*

**Proof** One could directly compute and see that $\tilde{\ell}$ is unbiased. But to give a little more insight into what motivates the definition of $\tilde{\ell}$, consider the conditions that unbiasedness imposes on it. We should have, for every $t$,

$$\mathbb{E}_{\tilde{y} \overset{\rho}{\sim} y}\left[\tilde{\ell}(t,\tilde{y})\right] = \ell(t,y)\ .$$

Considering the cases $y = +1$ and $y = -1$ separately, gives the equations

$$\begin{aligned}(1-\rho_{+1})\tilde{\ell}(t,+1) + \rho_{+1}\tilde{\ell}(t,-1) &= \ell(t,+1)\ ,\\ (1-\rho_{-1})\tilde{\ell}(t,-1) + \rho_{-1}\tilde{\ell}(t,+1) &= \ell(t,-1)\ .\end{aligned}$$

Solving these two equations for $\tilde{\ell}(t,+1)$ and $\tilde{\ell}(t,-1)$ gives

$$\tilde{\ell}(t,+1) = \frac{(1-\rho_{-1})\ell(t,+1) - \rho_{+1}\ell(t,-1)}{1-\rho_{+1}-\rho_{-1}} ,$$
$$\tilde{\ell}(t,-1) = \frac{(1-\rho_{+1})\ell(t,-1) - \rho_{-1}\ell(t,+1)}{1-\rho_{+1}-\rho_{-1}} .$$

The second part of the lemma follows by observing that $\ell_\alpha$ is bounded too. ■

In Section 3, we saw that in the noise-free case one can bound the deficit utility $\mathcal{U}^* - \mathcal{U}(f;D)$ by using a weighted surrogate loss approach with $\alpha = \delta_D^*$. In the presence of noisy labels, we can try to learn a good predictor that optimizes the measure $\mathcal{U}$ of the form (1) by minimizing the sample average

$$\hat{f} \leftarrow \operatorname*{argmin}_{f\in\mathcal{F}} \widehat{R}_{\tilde{\ell}_\alpha}(f) := \sum_{i=1}^{n} \tilde{\ell}_\alpha(f(X_i), \tilde{Y}_i) . \tag{3}$$

where $\alpha = \delta_D^*$ as before. By unbiasedness of $\tilde{\ell}_\alpha$ (Lemma 7), we know that, for any fixed $f \in \mathcal{F}$, the above sample average converges to $R_{\ell_\alpha,D}(f)$ even though the former is computed using noisy labels whereas the latter depends on the true labels. The following result gives a performance guarantee for this procedure in terms of the Rademacher complexity of the function class $\mathcal{F}$. The main idea in the proof is to use the contraction principle for Rademacher complexity to get rid of the dependence on the proxy loss $\tilde{\ell}_\alpha$. The price to pay for this is $L_\rho$, the Lipschitz constant of $\tilde{\ell}_\alpha$.

**Lemma 8** *Let $\ell(t,y)$ be $L$-Lipschitz in $t$ (for every $y$). Then, for any $\alpha \in (0,1)$, with probability at least $1-\delta$,*

$$\max_{f\in\mathcal{F}} |\widehat{R}_{\tilde{\ell}_\alpha}(f) - R_{\tilde{\ell}_\alpha,D_\rho}(f)| \le 2L_\rho \mathfrak{R}(\mathcal{F}) + \sqrt{\frac{\log(1/\delta)}{2n}}$$

*where $\mathfrak{R}(\mathcal{F}) := \mathbb{E}_{X_i,\epsilon_i}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\epsilon_i f(X_i)\right]$ is the Rademacher complexity of the function class $\mathcal{F}$ and $L_\rho \le 2L/(1-\rho_{+1}-\rho_{-1})$ is the Lipschitz constant of $\tilde{\ell}_\alpha$. Note that $\epsilon_i$'s are iid Rademacher (symmetric Bernoulli) random variables.*

**Proof** By the basic Rademacher bound on the maximal deviation between risks and empirical risks over $f \in \mathcal{F}$, we get

$$\max_{f\in\mathcal{F}} |\widehat{R}_{\tilde{\ell}_\alpha}(f) - R_{\tilde{\ell}_\alpha,D_\rho}(f)| \le 2 \cdot \mathfrak{R}(\tilde{\ell}_\alpha \circ \mathcal{F}) + \sqrt{\frac{\log(1/\delta)}{2n}}$$

where

$$\mathfrak{R}(\tilde{\ell}_\alpha \circ \mathcal{F}) := \mathbb{E}_{X_i,\tilde{Y}_i,\epsilon_i}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \epsilon_i \tilde{\ell}_\alpha(f(X_i),\tilde{Y}_i)\right]$$

If $\ell$ is $L$-Lipschitz then for any $\alpha \in (0,1)$, $\tilde{\ell}_\alpha$ is $L_\rho$ Lipschitz for $L_\rho = (1+|\rho_{+1}-\rho_{-1}|)L/(1-\rho_{+1}-\rho_{-1}) \leq 2L/(1-\rho_{+1}-\rho_{-1})$ and hence by the Lipschitz composition property of Rademacher averages, we have

$$\mathfrak{R}(\tilde{\ell}_\alpha \circ \mathcal{F}) \leq L_\rho \cdot \mathfrak{R}(\mathcal{F}) \ .$$

■

The above lemma immediately leads to a performance bound for $\hat{f}$ with respect to the clean distribution $D$. Our first main result is stated in the theorem below. The proof relies on using the $\alpha$-CC property of the modified surrogate loss function.

**Theorem 9** *For any $\alpha \in (0,1)$, with probability at least $1-\delta$,*

$$R_{\ell_\alpha,D}(\hat{f}) \leq \min_{f \in \mathcal{F}} R_{\ell_\alpha,D}(f) + 4L_\rho \mathfrak{R}(\mathcal{F}) + 2\sqrt{\frac{\log(1/\delta)}{2n}} \ .$$

*Furthermore, if $\ell_\alpha$ is $\alpha$-CC, then for the choice $\alpha = \delta_D^*$, there exists a nondecreasing function $\zeta_{\ell,\alpha}$ with $\zeta_{\ell,\alpha}(0) = 0$ such that,*

$$\mathcal{U}^* - \mathcal{U}(\hat{f}; D) \leq \zeta_{\ell,\alpha}\left( \min_{f \in \mathcal{F}} R_{\ell_\alpha,D}(f) - \min_f R_{\ell_\alpha,D}(f) + 4L_\rho \mathfrak{R}(\mathcal{F}) + 2\sqrt{\frac{\log(1/\delta)}{2n}} \right) \ .$$

**Proof** Let $f^\star$ be the minimizer of $R_{\ell_\alpha,D}(\cdot)$ over $\mathcal{F}$. We have

$$\begin{aligned}
& R_{\ell_\alpha,D}(\hat{f}) - R_{\ell_\alpha,D}(f^\star) \\
&= R_{\tilde{\ell}_\alpha,D_\rho}(\hat{f}) - R_{\tilde{\ell}_\alpha,D_\rho}(f^\star) \\
&= \widehat{R}_{\tilde{\ell}_\alpha}(\hat{f}) - \widehat{R}_{\tilde{\ell}_\alpha}(f^\star) + (R_{\tilde{\ell}_\alpha,D_\rho}(\hat{f}) - \widehat{R}_{\tilde{\ell}_\alpha}(\hat{f})) \\
&\quad + (\widehat{R}_{\tilde{\ell}_\alpha}(f^\star) - R_{\tilde{\ell}_\alpha,D_\rho}(f^\star)) \\
&\leq 0 + 2 \max_{f \in \mathcal{F}} \ |\widehat{R}_{\tilde{\ell}_\alpha}(f) - R_{\tilde{\ell}_\alpha,D_\rho}(f)| \ .
\end{aligned}$$

We can now apply Lemma 8 to control the last quantity above, and thus obtain the first statement of the theorem. Now, if $\ell_\alpha$ is $\alpha$-CC, then for $\alpha = \delta_D^*$, we know from Lemma 5 there exists a convex, invertible, nondecreasing transformation $\psi_\ell$ with $\psi_{\ell,\alpha}(0) = 0$ such that,

$$\psi_{\ell,\alpha}(R_\alpha(f) - R_\alpha^*) \leq R_{\ell_\alpha,D}(f) - \min_f R_{\ell_\alpha,D}(f)$$

Subtracting $\min_f R_{\ell_\alpha,D}(f)$ off either sides of the first inequality in the theorem statement, and realizing that $\psi_{\ell,\alpha}^{-1}$ is nondecreasing as well, with $\psi_{\ell,\alpha}^{-1}(0) = 0$, we get:

$$R_\alpha(\hat{f}) - R_\alpha^* \leq \psi_\ell^{-1}\left( \min_{f \in \mathcal{F}} R_{\ell,D}(f) - \min_f R_{\ell,D}(f) + 4L_\rho \mathfrak{R}(\mathcal{F}) + 2\sqrt{\frac{\log(1/\delta)}{2n}} \right) \ .$$

Finally we can use Lemma 3 to bound $\mathcal{U}^* - \mathcal{U}(\hat{f}; D)$, by setting $\zeta_{\ell,\alpha} = (a_{11} + a_{00} - a_{01} - a_{10})\psi_{\ell,\alpha}^{-1}$. ■

The term on the right hand side involves both approximation error (that is small if $\mathcal{F}$ is large) and estimation error (that is small if $\mathcal{F}$ is small). However, by appropriately increasing the richness of the class $\mathcal{F}$ with sample size, we can ensure that the utility of $\hat{f}$ approaches the optimal utility under the true distribution. This is despite the fact that the method of unbiased estimators computes the empirical minimizer $\hat{f}$ on a sample from the noisy distribution. Getting the optimal empirical minimizer $\hat{f}$ is efficient if $\tilde{\ell}_\alpha$, or rather $\tilde{\ell}$, is convex. Next, we address the issue of convexity of $\tilde{\ell}$.

### 4.1 Convex losses and their estimators

Note that the loss $\tilde{\ell}$ may not be convex even if we start with a convex $\ell$. An example is provided by the familiar hinge loss $\ell_{\text{hin}}(t, y) = [1 - yt]_+$. Stempfel and Ralaivola (2009) showed that $\tilde{\ell}_{\text{hin}}$ is not convex in general (of course, when $\rho_{+1} = \rho_{-1} = 0$, it is convex). Below we provide a simple condition to ensure convexity of $\tilde{\ell}$.

**Lemma 10** *Suppose $\ell(t, y)$ is convex and twice differentiable almost everywhere in $t$ (for every $y$) and also satisfies the symmetry property*

$$\forall t \in \mathbb{R},\ \ell''(t, y) = \ell''(t, -y)\ .$$

*Then $\tilde{\ell}(t, y)$ is also convex in $t$.*

**Proof** Let us compute $\tilde{\ell}''(t, y)$ (recall that differentiation is w.r.t. $t$) and show that it is non-negative under the symmetry condition $\ell''(t, y) = \ell''(t, -y)$. We have

$$\begin{aligned}
\tilde{\ell}''(t, y) &= \frac{(1 - \rho_{-y})\ell''(t, y) - \rho_y \ell''(t, -y)}{1 - \rho_{+1} - \rho_{-1}} \\
&= \frac{(1 - \rho_{-y})\ell''(t, y) - \rho_y \ell''(t, y)}{1 - \rho_{+1} - \rho_{-1}} \\
&= \frac{(1 - \rho_{-y} - \rho_y)\ell''(t, y)}{1 - \rho_{+1} - \rho_{-1}} \\
&= \ell''(t, y) \geq 0\ ,
\end{aligned}$$

since $\ell$ is convex in $t$. ■

Examples satisfying the conditions of the lemma above are the squared loss $\ell_{\text{sq}}(t, y) = (t - y)^2$, the logistic loss $\ell_{\log}(t, y) = \log(1 + \exp(-ty))$ and the Huber loss:

$$\ell_{\text{Hub}}(t, y) = \begin{cases} -4yt & \text{if } yt < -1 \\ (t - y)^2 & \text{if } -1 \leq yt \leq 1 \\ 0 & \text{if } yt > 1 \end{cases}$$

Consider the case where $\tilde{\ell}$ turns out to be non-convex when $\ell$ is convex, as in $\tilde{\ell}_{\text{hin}}$. In the online learning setting (where the adversary chooses a sequence of examples, and the prediction of a learner at round $i$ is based on the history of $i-1$ examples with independently

flipped labels) which we will discuss shortly, we would use a stochastic mirror descent type algorithm (Nemirovski et al., 2009) to arrive at risk bounds similar to Theorem 9. Then, we only need the expected loss to be convex and therefore $\ell_{\text{hin}}$ does not present a problem. At first blush, it may appear that we do not have much hope of obtaining $\hat{f}$ in the iid setting efficiently. However, Lemma 8 provides a clue.

We will now focus on the function class $\mathcal{W}$ of hyperplanes. Even though $\widehat{R}_{\tilde{\ell}}(\mathbf{w})$ is non-convex, it is uniformly close to $R_{\tilde{\ell},D_\rho}(\mathbf{w})$. Since $R_{\tilde{\ell},D_\rho}(\mathbf{w}) = R_{\ell,D}(\mathbf{w})$, this shows that $\widehat{R}_{\tilde{\ell}}(\mathbf{w})$ is uniformly close to a convex function over $\mathbf{w} \in \mathcal{W}$. The following result shows that we can therefore approximately minimize $F(\mathbf{w}) = \widehat{R}_{\tilde{\ell}}(\mathbf{w})$ by minimizing the biconjugate $F^{\star\star}$. Recall that the (Fenchel) biconjugate $F^{\star\star}$ is the largest convex function that minorizes $F$.

**Lemma 11** *Let $F : \mathcal{W} \to \mathbb{R}$ be a non-convex function defined on function class $\mathcal{W}$ such it is $\varepsilon$-close to a convex function $G : \mathcal{W} \to \mathbb{R}$:*

$$\forall \mathbf{w} \in \mathcal{W},\ |F(\mathbf{w}) - G(\mathbf{w})| \leq \varepsilon$$

*Then any minimizer of $F^{\star\star}$ is a $2\varepsilon$-approximate (global) minimizer of $F$.*

**Proof** Since $F \geq G - \varepsilon$ and $F^{\star\star}$ is the largest convex function that minorizes $F$, we must have $F^{\star\star} \geq G - \varepsilon$. This means that $F^{\star\star} + 2\varepsilon \geq G + \varepsilon \geq F$. Thus, $F$ is sandwiched between $F^{\star\star} + 2\varepsilon$ and $F^{\star\star}$. The lemma follows directly from this. ■

Now, the following theorem establishes bounds for the case when $\tilde{\ell}$ is non-convex, via the solution obtained by minimizing the convex function $F^{**}$.

**Theorem 12** *Let $\ell$ be a loss, such as the hinge loss, for which $\tilde{\ell}$ is non-convex. Let $\mathcal{W} = \{\mathbf{w} : \|\mathbf{w}_2\| \leq W_2\}$, let $\|X_i\|_2 \leq X_2$ almost surely, and let $\hat{\mathbf{w}}_{\text{approx}}$ be any (exact) minimizer of the convex problem*

$$\min_{\mathbf{w}\in\mathcal{W}}\ F^{\star\star}(\mathbf{w})\ ,$$

*where $F^{\star\star}(\mathbf{w})$ is the (Fenchel) biconjugate of the function $F(\mathbf{w}) = \widehat{R}_{\tilde{\ell}_\alpha}(\mathbf{w})$, where $\alpha = \delta_D^*$. Then, with probability at least $1-\delta$, $\hat{\mathbf{w}}_{\text{approx}}$ is a $2\varepsilon$-minimizer of $\widehat{R}_{\tilde{\ell}_\alpha}(\cdot)$ where*

$$\varepsilon = \frac{2L_\rho X_2 W_2}{\sqrt{n}} + \sqrt{\frac{\log(1/\delta)}{2n}}\ .$$

*Therefore, with probability at least $1-\delta$,*

$$R_{\ell_\alpha,D}(\hat{\mathbf{w}}_{\text{approx}}) \leq \min_{\mathbf{w}\in\mathcal{W}} R_{\ell_\alpha,D}(\mathbf{w}) + 4\varepsilon\ .$$

**Proof** The first part of the theorem follows by combining Lemma 8 and Lemma 11, using the fact that if $\|\mathbf{w}\|_2 \leq W_2$ for any $\mathbf{w}$ and $\|X_i\|_2 \leq X_2$ then, $\mathfrak{R}(\mathcal{W}) \leq W_2 X_2/\sqrt{n}$. Note that Theorem 9 is true also for $2\varepsilon$-minimizers of the empirical risk $\widehat{R}_{\tilde{\ell}_\alpha}$ provided we add $2\varepsilon$ to the right hand side. ■

Numerical or symbolic computation of the biconjugate of a multidimensional function is difficult, in general, but can be done in special cases. It will be interesting to see if techniques from Computational Convex Analysis (Lucet, 2010) can be used to efficiently compute the biconjugate above.

### 4.2 Online learning setting

Consider the setting where an adversary chooses a sequence $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ of examples. At time $i$, the learner has to make a prediction based on $(\mathbf{x}_1, \tilde{y}_1), \ldots, (\mathbf{x}_{i-1}, \tilde{y}_{i-1})$ and $\mathbf{x}_i$. But the learner's cumulative loss as well as that of the best fixed predictor in hindsight are both computed using the true labels $y_i$. Note that if $\ell(t, y)$ is convex in $t$ (for every $y$), and for a given $\alpha \in (0, 1)$ we choose $\lambda_1 \in \partial \ell_\alpha(t, y)$ and $\lambda_2 \in \partial \ell_\alpha(t, -y)$, (where $\partial \ell_\alpha$ is the subdifferential w.r.t. $t$) we have

$$\mathbb{E}_{\tilde{y}} \left[ g(t, \tilde{y}) \right] \in \partial \ell_\alpha(t, y) \tag{4}$$

where

$$g(t, y) = \frac{(1 - \rho_{-y})\lambda_1 - \rho_y \lambda_2}{1 - \rho_{+1} - \rho_{-1}} \tag{5}$$

We show that Algorithm 1 indeed satisfies low regret (in expectation) on the original sequence chosen by the adversary even though it only receives noisy versions of the labels. We fix the function class to be the set $\mathcal{W}$ of bounded-norm hyperplanes.

**Algorithm 1:** Online learning using unbiased gradients

Choose learning rate $\gamma > 0$
$\mathcal{W} = \{\mathbf{w} : \|\mathbf{w}\|_2 \leq W_2\}$
$\Pi_{\mathcal{W}}(\cdot) =$ Euclidean projection onto $\mathcal{W}$
Initialize $\mathbf{w}_0 \leftarrow \mathbf{0}$
**for** $i = 1$ to $n$ **do**
  Receive $\mathbf{x}_i \in \mathbb{R}^d$
  Predict $\langle \mathbf{w}_{i-1}, \mathbf{x}_i \rangle$
  Receive noisy label $\tilde{y}_i$
  Update $\mathbf{w}_i \leftarrow \Pi_{\mathcal{W}} \left( \mathbf{w}_{i-1} - \gamma g(\langle \mathbf{w}_{i-1}, \mathbf{x}_i \rangle, \tilde{y}_i) \mathbf{x}_i \right)$ where $g(\cdot, \cdot)$ is defined in (5)
**end for**

**Theorem 13** *Let $\ell(t, y)$ be convex and L-Lipschitz in $t$ (for every $y$). Fix an arbitrary sequence $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$, and $\alpha \in (0, 1)$. If Algorithm 1 is run on noisy data set $(\mathbf{x}_1, \tilde{y}_1), \ldots, (\mathbf{x}_n, \tilde{y}_n)$ with learning rate $\gamma = W_2/(X_2 L_\rho \sqrt{n})$ where $\tilde{y}_i$ is noisy version of $y_i$ with noise rates $\rho_{+1}, \rho_{-1}$, then we have*

$$\mathbb{E}_{\tilde{y}_{1:n}} \left[ \sum_{i=1}^{n} \ell_\alpha(\langle \mathbf{w}_{i-1}, \mathbf{x}_i \rangle, y_i) \right] - \min_{\|\mathbf{w}\|_2 \leq W_2} \sum_{i=1}^{n} \ell_\alpha(\langle \mathbf{w}, \mathbf{x}_i \rangle, y_i) \leq L_\rho X_2 W_2 \sqrt{n} \ ,$$

*where $L_\rho := (1 + |\rho_{+1} - \rho_{-1}|)L/(1 - \rho_{+1} - \rho_{-1})$ and it is assumed that $\|\mathbf{x}_i\| \leq X_2$ for all $i \in [n]$.*

**Proof** Let us use the abbreviation $g_i$ for $g(\langle \mathbf{w}_{i-1}, \mathbf{x}_i \rangle, \tilde{y}_i)\mathbf{x}_i$ so that the update in Algorithm 1 becomes $\mathbf{w}_i \leftarrow \Pi_{\mathcal{W}}(\mathbf{w}_{i-1} - \gamma g_i)$. It is well known (Zinkevich, 2003) that, for any $\mathbf{w}$,

$$\sum_{i=1}^{n} \langle g_i, \mathbf{w}_{i-1} - \mathbf{w} \rangle \leq \frac{\gamma}{2} \sum_{i=1}^{n} \|g_i\|^2 + \frac{\|\mathbf{w}\|^2}{2\gamma} \ . \tag{6}$$

Since $\ell$ is $L$-Lipschitz, the $\lambda_1, \lambda_2$ appearing in the definition (5) of $g(\cdot,\cdot)$ satisfy $|\lambda_1|, |\lambda_2| \leq L$. This implies $|g(t,y)| \leq (1+|\rho_{+1}-\rho_{-1}|)L/(1-\rho_{+1}-\rho_{-1}) = L_\rho$ and hence $\|g_i\| \leq L_\rho X_2$. Thus, we have, for any $\mathbf{w}$ with $\|\mathbf{w}\| \leq W_2$, $\sum_{i=1}^n \langle g_i, \mathbf{w}_{i-1} - \mathbf{w}\rangle \leq \frac{\gamma L_\rho^2 X_2^2 n}{2} + \frac{W_2^2}{2\gamma}$. Choosing $\gamma = (W_2/L_\rho X_2)\frac{1}{\sqrt{n}}$, we get $\sum_{i=1}^n \langle g_i, \mathbf{w}_{i-1} - \mathbf{w}\rangle \leq L_\rho X_2 W_2 \sqrt{n}$. Note that $\mathbf{w}_{i-1}$ only depends on $\tilde{y}_{1:i-1}$. Hence

$$\mathbb{E}_{\tilde{y}_i}\left[\langle g_i, \mathbf{w}_{i-1} - \mathbf{w}\rangle \mid \tilde{y}_{1:i-1}\right] = \left\langle \mathbb{E}_{\tilde{y}_i}\left[g_i \mid \tilde{y}_{1:i-1}\right], \mathbf{w}_{i-1} - \mathbf{w}\right\rangle \geq \ell_\alpha(\langle \mathbf{w}_{i-1}, \mathbf{x}_i\rangle, y_i) - \ell_\alpha(\langle \mathbf{w}, \mathbf{x}_i\rangle, y_i)$$

because $\mathbb{E}_{\tilde{y}_i}\left[g_i \mid \tilde{y}_{1:i-1}\right] \in \partial_{\mathbf{w}=\mathbf{w}_{i-1}} \ell_\alpha(\langle \mathbf{w}, \mathbf{x}_i\rangle, y_i)$ by (4) and the chain rule for differentiation, and $\ell_\alpha(\langle \mathbf{w}, \mathbf{x}_i\rangle, y_i)$ is convex in $\mathbf{w}$. Thus, for any $\mathbf{w}$ with $\|\mathbf{w}\|_2 \leq W_2$,

$$\mathbb{E}_{\tilde{y}_{1:n}}\left[\sum_{i=1}^n \ell_\alpha(\langle \mathbf{w}_{i-1}, \mathbf{x}_i\rangle, y_i)\right] - \sum_{i=1}^n \ell_\alpha(\langle \mathbf{w}, \mathbf{x}_i\rangle, y_i) \leq L_\rho X_2 W_2 \sqrt{n}.$$

Since the above inequality is true for any $\mathbf{w}$ with $\|\mathbf{w}\|_2 \leq 1$, the statement of the theorem follows. ■

## 5. Approach of Surrogates for Weighted 0-1 Loss

The second approach is based on directly obtaining weighted surrogates for $\mathcal{U}$. We develop the method of "label-dependent" costs from two key observations. First, the Bayes classifier under noisy distribution, denoted by $\tilde{f}^*$, simply uses a threshold that, in general, is different from that under clean distribution. Second, $\tilde{f}^*$ is the minimizer of a certain weighted 0-1 loss under the noisy distribution. The framework we develop here generalizes known results for the uniform noise rate setting $\rho_{+1} = \rho_{-1}$ and offers a more fundamental insight into the problem.

From Lemma 2, we know that the optimal Bayes classifier $\tilde{f}^*$ under $D_\rho$ thresholds $\tilde{\eta}(X) = P(\tilde{Y} = 1|X)$ at a certain $\tilde{\delta}^*$. Now, noting that:

$$\tilde{\eta}(x) = (1-\rho_{+1})\eta(x) + \rho_{-1}(1-\eta(x)) = (1-\rho_{+1}-\rho_{-1})\eta(x) + \rho_{-1},$$

we see that $\tilde{f}^*$ can be written as:

$$\tilde{f}^*(x) = \text{sign}\left(\eta(x) - \frac{\tilde{\delta}^* - \rho_{-1}}{1-\rho_{+1}-\rho_{-1}}\right). \tag{7}$$

Not surprisingly, this optimal threshold simplifies for cost-sensitive performance measures. In particular, as shown in the following corollary, the optimal threshold for the AM measure does *not* change under the noisy distribution.

**Corollary 14** *1. For the Accuracy measure:*

$$\arg\max_f \mathcal{U}_{Acc}(f; D_\rho) = \arg\min_f R_{D_\rho}(f) = \text{sign}\left(\eta(x) - \frac{1/2 - \rho_{-1}}{1-\rho_{+1}-\rho_{-1}}\right).$$

*2. For the AM Measure:*

$$\arg\max_f \mathcal{U}_{AM}(f; D) = \arg\max_f \mathcal{U}_{AM}(f; D_\rho) = \text{sign}(\eta(x) - \pi).$$

**Proof**

1. For the 0-1 loss, $\delta^*_D = \tilde{\delta}^* = 1/2$ and the result is immediate.
2. We know $\delta^*_D = \pi$ from Lemma 2. Also, $\tilde{\delta}^* = P(\tilde{Y} = 1) = (1 - \rho_{-1} - \rho_{+1})\pi + \rho_{-1}$. Substituting in (7), we observe that the threshold remains $\pi$.

■

Interestingly, this *noisy* Bayes classifier can also be obtained as the minimizer of a weighted 0-1 loss; which as we will show, allows us to "correct" for the threshold under the noisy distribution. Let us first introduce the notion of label-dependent costs for binary classification. We can write the 0-1 loss as a label-dependent loss as follows:

$$1_{\{\text{sign}(f(X))\neq Y\}} = 1_{\{Y=1\}}1_{\{f(X)\leq 0\}} + 1_{\{Y=-1\}}1_{\{f(X)>0\}}$$

Clearly, the classical 0-1 loss is *unweighted*. Consider the $\alpha$-weighted 0-1 loss (which is a special case of the weighted loss (2)):

$$U_\alpha(t, y) = (1 - \alpha)1_{\{y=1\}}1_{\{t\leq 0\}} + \alpha 1_{\{y=-1\}}1_{\{t>0\}},$$

where $\alpha \in (0, 1)$. In fact we see that minimization w.r.t. the 0-1 loss is equivalent to that w.r.t. $U_{1/2}(f(X), Y)$. It is not a coincidence that Bayes optimal $f^*$ has a threshold $1/2$. The following lemma (Scott, 2012) shows that in fact for any $\alpha$-weighted 0-1 loss, the minimizer thresholds $\eta(x)$ at $\alpha$.

**Lemma 15 ($\alpha$-weighted Bayes optimal (Scott, 2012))** *For $\alpha \in (0, 1)$,*

$$f^*_\alpha := \arg\min_f R_\alpha(f) = \text{sign}(\eta(x) - \alpha) \ .$$

At this juncture, we are interested in the following question: For a given $\delta$, does there exist an $\alpha \in (0, 1)$ such that the minimizer of $U_\alpha$-risk under noisy distribution $D_\rho$ has the same sign as that of the Bayes optimal $f^*_\delta$? We now present our second main result in the following theorem that makes a stronger statement — the $U_\alpha$-risk under noisy distribution $D_\rho$ is linearly related to $U_\delta$-risk under the clean distribution $D$. The corollary of the theorem answers the question in the affirmative.

**Theorem 16** *For any given $\delta \in (0, 1)$, for the choices,*

$$\alpha^* = \rho_{-1} + (1 - \rho_{+1} - \rho_{-1})\delta \text{ and } A_\rho = 1 - \rho_{+1} - \rho_{-1},$$

*there exists a constant $B_X$ that is independent of $f$ such that, for all functions $f$,*

$$R_{\alpha^*, D_\rho}(f) = A_\rho R_{\delta, D}(f) + B_X.$$

**Proof** For simplicity, let us think of $f$ as $\{\pm 1\}$-valued. We have,

$$C_{\delta,D}(f) = \mathbb{E}_Y \left[ (1-\delta) 1_{\{Y=1\}} 1_{\{f(X)\neq 1\}} + \delta 1_{\{Y=-1\}} 1_{\{f(X)\neq -1\}} \right]$$

and

$$C_{\alpha,D_\rho}(f) = \mathbb{E}_{\tilde{Y}} \left[ (1-\alpha) 1_{\{\tilde{Y}=1\}} 1_{\{f(X)\neq 1\}} + \alpha 1_{\{\tilde{Y}=-1\}} 1_{\{f(X)\neq -1\}} \right].$$

Note that $R_{\delta,D}(f) = \mathbb{E}_X \left[ C_{\delta,D}(f) \right]$, and $R_{\alpha,D_\rho}(f) = \mathbb{E}_X \left[ C_{\alpha,D_\rho}(f) \right]$. Also note that $C_{\delta,D}(f) = (1-\delta)\eta(X)$ if $f(X) = -1$, and $C_{\delta,D}(f) = \delta(1-\eta(X))$ otherwise.
Similarly, $C_{\alpha,D_\rho}(f) = (1-\alpha)\tilde{\eta}(X)$ if $f(X) = -1$ and $C_{\alpha,D_\rho}(f) = \alpha(1-\tilde{\eta}(X))$ otherwise.
We want to find $A$ and $B$ such that the following equations hold simultaneously:

$$(1-\alpha)\tilde{\eta}(X) = A(1-\delta)\eta(X) + B$$
$$\alpha(1-\tilde{\eta}(X)) = A\delta(1-\eta(X)) + B$$

Using the relation between $\eta(X)$ and $\tilde{\eta}(X)$ and solving for $A$ we get,

$$A = \frac{(1-\rho_{+1}-\rho_{-1})\eta(X) + \rho_{-1} - \alpha}{\eta(X) - \delta}.$$

Choosing $\alpha = \alpha^* = \rho_{-1} + (1-\rho_{+1}-\rho_{-1})\delta$, and simplifying, we get a constant $A$ that depends only on the noise rates:

$$A = A_\rho = 1 - \rho_{+1} - \rho_{-1}.$$

Consequently,

$$B = \rho_{-1}(1-\alpha^*) + (\delta - \alpha^*)(1-\rho_{+1}-\rho_{-1})\eta(X).$$

Taking expectation with respect to $X$, we conclude:

$$R_{\alpha^*,D_\rho}(f) = A_\rho R_{\delta,D}(f) + B_X,$$

where $B_X = \mathbb{E}_X [B]$. ■

**Corollary 17** *Let $\alpha^* = \rho_{-1} + (1-\rho_{+1}-\rho_{-1})\delta_D^*$. The $\alpha^\star$-weighted Bayes optimal classifier under noisy distribution coincides with that of $\mathcal{U}$ measure under clean distribution:*

$$\operatorname*{argmin}_f R_{\alpha^*,D_\rho}(f) = \operatorname*{argmin}_f R_{\delta_D^*,D}(f) = \operatorname*{argmin}_f \mathcal{U}(f;D).$$

We are now ready to state our next main result — a certain weighted ERM is consistent: i.e. the "true" performance of the empirical minimizer w.r.t. the noisy distribution converges to the optimal performance $\mathcal{U}^*$ at a steady rate. The resulting bound has a striking resemblance to that of our first result in Theorem 9. The proof technique is similar to that of Theorem 9, and crucially relies on using the relationship established in Theorem 16.

**Theorem 18** *Given a convex loss function $\ell : \mathbb{R} \to [0, \infty)$ with Lipschitz constant $L$ such that it is classification-calibrated (i.e. $\ell^{'}(0) < 0$), consider the empirical risk minimization problem with noisy labels:*

$$\hat{f_\alpha} = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \ell_\alpha(f(X_i), \tilde{Y_i}). \tag{8}$$

*where $\ell_\alpha$ is defined as in (2). Then, for the choice of $\alpha^*$ in Corollary 17, there exists a nondecreasing function $\zeta_{\ell_{\alpha^\star}}$ with $\zeta_{\ell_{\alpha^\star}}(0) = 0$, such that the following bound holds with probability at least $1 - \delta$:*

$$\mathcal{U}^* - \mathcal{U}(\hat{f}_{\alpha^*}; D) \leq \frac{1}{A_\rho} \zeta_{\ell_{\alpha^\star}} \left( \min_{f \in \mathcal{F}} R_{\alpha^*, D_\rho}(f) - \min_f R_{\alpha^*, D_\rho}(f) + 4L\mathfrak{R}(\mathcal{F}) + 2\sqrt{\frac{\log(1/\delta)}{2n}} \right),$$

*where $A_\rho = 1 - \rho_{+1} - \rho_{-1}$.*

**Proof** From Corollary 4.1 of Scott (2012), we can infer that $\ell_\alpha$ is $\alpha$-CC for given $\alpha \in (0, 1)$, as $\ell$ is convex, classification-calibrated and $\ell^{'}(0) < 0$. Then, from Theorem 3.1 of Scott (2012), there exists an *invertible*, *non-decreasing* convex transformation $\psi_{\ell_\alpha}$ with $\psi_{\ell_\alpha}(0) = 0$ such that, for any $f$ and any distribution $D$,

$$\psi_{\ell_\alpha}(R_{\alpha, D}(f) - \min_f R_{\alpha, D}(f)) \leq R_{\ell_\alpha, D}(f) - \min_f R_{\ell_\alpha, D}(f).$$

Fix distribution to be $D_\rho$, and let $f = \hat{f}_\alpha$. The RHS of the above inequality can then be controlled similarly as in the proof of Theorem 9. It is easy to see that the Lipschitz constant of $\ell_\alpha$ is same as that of $\ell$, denoted $L$. With probability at least $1 - \delta$:

$$R_{\ell_\alpha, D_\rho}(\hat{f}_\alpha) - \min_{f \in \mathcal{F}} R_{\ell_\alpha, D_\rho}(f) \leq 4L\mathfrak{R}(\mathcal{F}) + 2\sqrt{\frac{\log(1/\delta)}{2n}}.$$

Now consider $R_{\alpha, D_\rho}(f) - \min_f R_{\alpha, D_\rho}(f)$. Using the linear relationship between $R_{\alpha, D_\rho}$ and $R_{\delta^*_D, D}$ at $\alpha^*$ (Theorem 16), we get $R_{\alpha^*, D_\rho}(f) - \min_f R_{\alpha^*, D_\rho}(f) = A_\rho(R_{\delta^*_D, D}(f) - R^*_{\delta^*_D})$. $B_X$ vanishes because it is constant for the distribution $D_\rho$. Note that $\psi^{-1}_{\ell_{\alpha^*}}$ is nondecreasing as well and $\psi^{-1}_{\ell_{\alpha^*}}(0) = 0$. Subtracting $\min_f R_{\alpha^*, D_\rho}(f)$ from both sides of the second inequality above, we get: With probability at least $1 - \delta$,

$$R_{\delta^*_D, D}(\hat{f}_{\alpha^*}) - R^*_{\delta^*_D} \leq A_\rho^{-1} \psi^{-1}_{\ell_{\alpha^\star}} \left( \min_{f \in \mathcal{F}} R_{\alpha^*, D_\rho}(f) - \min_f R_{\alpha^*, D_\rho}(f) + 4L\mathfrak{R}(\mathcal{F}) + 2\sqrt{\frac{\log(1/\delta)}{2n}} \right).$$

Invoking Lemma 3 and setting $\zeta_{\ell_{\alpha^\star}} = (a_{11} + a_{00} - a_{10} - a_{01})\psi^{-1}_{\ell_{\alpha^\star}}$, the proof is complete. ■

## 6. Experiments

In our first set of experiments, we demonstrate the robustness of the proposed algorithms to increasing rates of label noise on synthetic and real-world data sets. In our second set of

experiments, we also conduct a comparison of the performance of our two proposed methods with state-of-the-art methods for dealing with random label noise. In our experiments, we use the two utility measures listed in Proposition 1, i.e. $\mathcal{U}_{Acc}$ and $\mathcal{U}_{AM}$; note that the utility measures are computed with respect to the clean distribution. For given noise rates $\rho_{+1}$ and $\rho_{-1}$, labels are flipped accordingly. To account for randomness in the flips to simulate a given noise rate, we repeat each experiment 3 times, with independent corruptions of the data set for same setting of $\rho_{+1}$ and $\rho_{-1}$, and present the mean accuracy over the trials. Specifically, we divide each data set randomly into three training and test sets, and compute average utility over 3 train-test splits. We use cross-validation to tune parameters specific to the algorithms. Note that we perform cross-validation on a separate validation set with *noisy* labels. In our final set of experiments, we address a practical question of specifying true noise rates to the algorithms, and study how misspecification of noise rates affects the performance of the algorithms.

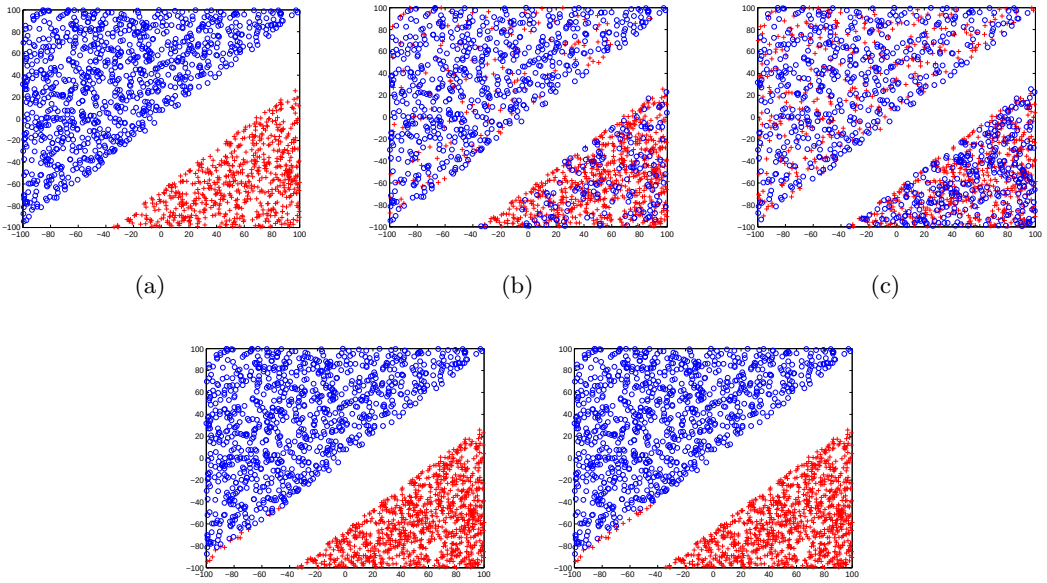
*Proposed methods.* For evaluation, we choose the following representative algorithms based on each of the two proposed methods: For the method in Section 4, we use unbiased estimator of the logistic loss. Here, the resulting ERM, i.e. (3) with $\ell_{\log}$, is solved using a gradient descent procedure. We refer to this as $\tilde{\ell}_{\log}$ for ease in the remainder of the section. For the method in Section 5 we use the widely-used C-SVM (Liu et al., 2003; Mordelet and Vert, 2014) method as well as weighted logistic regression, wherein we apply different costs on positive and negative examples in the respective loss functions. We use the `libsvm` library to solve the resulting ERM problems, i.e. (8) with $\ell_{\text{hin}}$ or $\ell_{\log}$ respectively. In all the cases, we tune the parameters $\alpha$, $\rho_{+1}$ and $\rho_{-1}$ by cross-validation (on noisy validation set).

### 6.1 Synthetic data

First, we use the synthetic 2D linearly separable data set shown in Figure 1(a). We observe from experiments that our methods achieve over 90% accuracy even when $\rho_{+1} = \rho_{-1} = 0.4$. Figure 1 shows the performance of $\tilde{\ell}_{\log}$ on the data set for different noise rates. Next, we use a 2D UCI benchmark non-separable data set ('banana'). The data set and classification results using C-SVM (which corresponds to vanilla SVM for uniform noise rates, $\alpha^* = 1/2$) are shown in Figure 2. The results for higher noise rates are impressive as observed from Figures 2(d) and 2(e). The 'banana' data set has been used in previous research on classification with noisy labels. In particular, the Random Projection classifier (Stempfel and Ralaivola, 2007) that learns a kernel perceptron in the presence of noisy labels achieves about 84% accuracy at $\rho_{+1} = \rho_{-1} = 0.3$ as observed from our experiments (as well as shown by Stempfel and Ralaivola, 2007), and the random hyperplane sampling method (Stempfel et al., 2007) gets about the same accuracy at $(\rho_{+1}, \rho_{-1}) = (0.2, 0.4)$ (as reported by Stempfel et al., 2007). Contrast these with C-SVM that achieves about 90% accuracy at $\rho_{+1} = \rho_{-1} = 0.2$ and over 88% accuracy at $\rho_{+1} = \rho_{-1} = 0.4$.

### 6.2 Comparison with state-of-the-art methods on UCI benchmark data

We next compare our methods with three state-of-the-art methods for dealing with random classification noise: Random Projection (RP) classifier (Stempfel and Ralaivola, 2007), NHERD (Crammer and Lee, 2010) (*project* and *exact* variants, which were shown to be the



(a) (b) (c)

(d) (e)

Figure 1: Classification of linearly separable synthetic data set using $\tilde{\ell}_{\log}$. The noise-free data is shown in the leftmost panel. Plots (b) and (c) show training data corrupted with noise rates ($\rho_{+1} = \rho_{-1} = \rho$) 0.2 and 0.4 respectively. Plots (d) and (e) show the corresponding classification results. The algorithm achieves 98.5% accuracy even at 0.4 noise rate per class. (Best viewed in color).

best performing variants from among the NHERD family of methods proposed by Crammer et al. 2006, 2009; Dredze et al. 2008), and perceptron algorithm with margin (PAM) which was shown to be robust to label noise by Khardon and Wachman (2007). We use seven standard UCI classification data sets listed in Table 1; here, data sets 1 through 6 are preprocessed and made available by Gunnar Rätsch.[1]

*Using linear kernel.* Results for the accuracy measure, for different settings of noise rates, using linear kernel in the compared methods, are shown in Table 2. C-SVM is competitive in 5 out of 7 data sets (Breast cancer, Thyroid, German, Image and Spambase), while relatively poorer in the other two. Note that in many cases, especially when $\rho_{+1} = \rho_{-1}$, the standard SVM (i.e. where positive and negative examples are weighted equally) as well as C-SVM (where $\alpha$ parameter that controls the relative weighting is tuned) yield the same accuracy, indicating that the cross-validation effectively selects equal weights; recall that the theory indeed suggests when the noise rates are equal, the optimal choice of weights are equal, i.e. $\alpha = 1/2$ (see Section 5). The corresponding results for the AM measure, for different settings of noise rates, are shown in Table 5. We find that C-SVM is competitive in three data sets, and NHERD is competitive in most of the data sets. Also observe that

1. `http://theoval.cmp.uea.ac.uk/matlab`

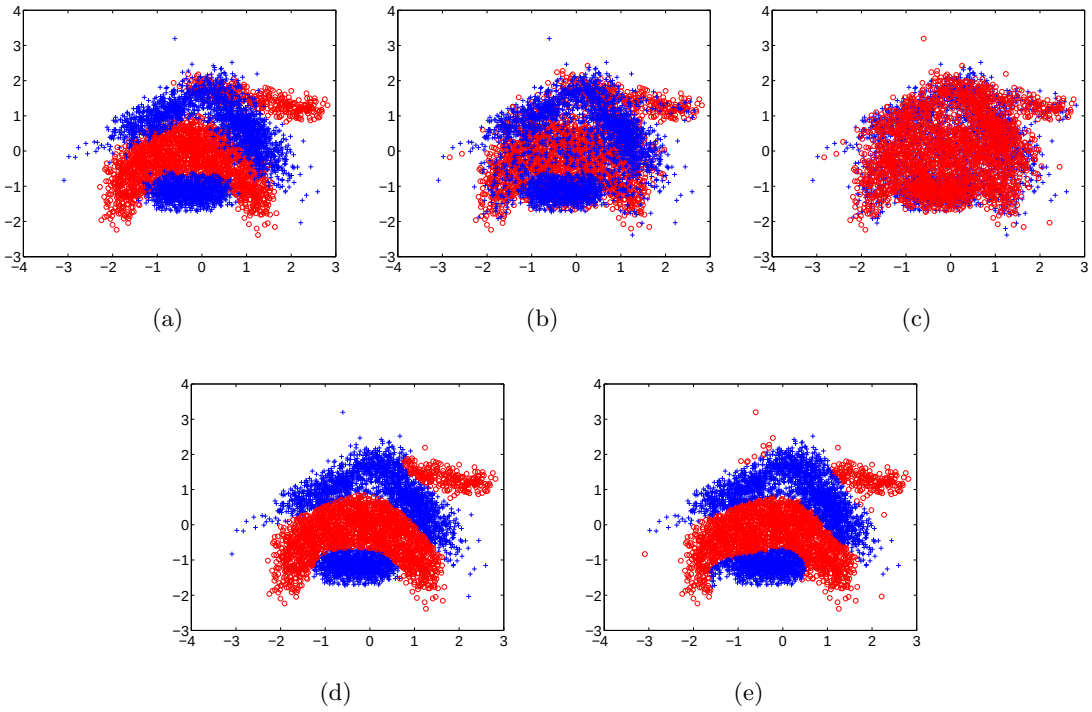


Figure 2: Classification of 'banana' data set using C-SVM. The noise-free data is shown in (a). Plots (b) and (c) show training data corrupted with noise rates ($\rho_{+1} = \rho_{-1} = \rho$) 0.2 and 0.4 respectively. Note that for $\rho_{+1} = \rho_{-1}$, $\alpha^* = 1/2$ (i.e. C-SVM reduces to regular SVM). Plots (d) and (e) show the corresponding classification results (Accuracies are 90.6% and 88.5% respectively). Even when 40% of the labels are corrupted ($\rho_{+1} = \rho_{-1} = 0.4$), the algorithm recovers the class structures as observed from plot (e). Note that the accuracy of the method at $\rho = 0$ is 90.8%.

at high noise rates AM is a more reliable measure of performance — in case of the four data sets Breast cancer, Diabetes, Thyroid and German which have class imbalance, the classifier optimized for the accuracy measure (Table 2) tends to bias its predictions towards the majority class (suggested by accuracy values matching the class imbalance ratio) but the achieved AM values are low.

We present the results for logistic loss based methods, using linear kernel, in Tables 3 and 6. As in the case of SVM based methods, we find, when $\rho_{+1} = \rho_{-1}$, the standard logistic regression (i.e. where positive and negative examples are weighted equally) as well as weighted logistic regression in the third column (where $\alpha$ parameter that controls the relative weighting is tuned) yield the same accuracy, indicating that the cross-validation effectively selects equal weights. In terms of accuracy, we find that $\tilde{\ell}_{\log}$ (in the second column) is competitive in all the data sets, wherease in terms of AM measure (see Table 6), (weighted) logistic regression performs the best more often.

| Data set | Dim | Num. Positives | Num. Negatives |
|---|---|---|---|
| Breast cancer | 9 | 77 | 186 |
| Diabetes | 8 | 268 | 500 |
| Thyroid | 5 | 65 | 150 |
| German | 20 | 300 | 700 |
| Heart | 13 | 120 | 150 |
| Image | 18 | 1188 | 898 |
| Spambase | 57 | 1813 | 2788 |

Table 1: UCI data sets used in experiments.

| Data set | Noise rates | SVM | C-SVM | PAM | NHERD | RP |
|---|---|---|---|---|---|---|
| Breast cancer | $\rho_{+1} = \rho_{-1} = 0.2$ | **70.25** | **70.25** | 68.42 | 64.90 | 38.95 |
| | $\rho_{+1} = 0.3, \rho_{-1} = 0.1$ | **71.53** | **71.53** | 69.97 | 65.68 | 66.93 |
| | $\rho_{+1} = \rho_{-1} = 0.4$ | 69.49 | **70.77** | 44.25 | 56.50 | 55.95 |
| Diabetes | $\rho_{+1} = \rho_{-1} = 0.2$ | **75.35** | **75.35** | 62.76 | 73.18 | 71.09 |
| | $\rho_{+1} = 0.3, \rho_{-1} = 0.1$ | **75.52** | **75.52** | 57.64 | **74.74** | 73.26 |
| | $\rho_{+1} = \rho_{-1} = 0.4$ | 68.84 | 68.84 | 51.52 | **71.09** | 68.58 |
| Thyroid | $\rho_{+1} = \rho_{-1} = 0.2$ | **81.94** | **81.94** | 63.58 | 78.49 | 78.89 |
| | $\rho_{+1} = 0.3, \rho_{-1} = 0.1$ | **86.63** | **86.63** | 45.48 | **87.78** | 78.89 |
| | $\rho_{+1} = \rho_{-1} = 0.4$ | 76.63 | 76.63 | 70.98 | **85.95** | 70.69 |
| German | $\rho_{+1} = \rho_{-1} = 0.2$ | **72.87** | **72.87** | 55.47 | 67.80 | 67.72 |
| | $\rho_{+1} = 0.3, \rho_{-1} = 0.1$ | **69.46** | **69.46** | 50.02 | 67.80 | 63.93 |
| | $\rho_{+1} = \rho_{-1} = 0.4$ | **62.20** | 56.60 | 41.33 | 54.80 | 56.27 |
| Heart | $\rho_{+1} = \rho_{-1} = 0.2$ | **82.96** | **82.96** | 73.09 | **82.96** | 76.05 |
| | $\rho_{+1} = 0.3, \rho_{-1} = 0.1$ | 77.53 | 77.53 | 71.60 | **81.48** | 78.77 |
| | $\rho_{+1} = \rho_{-1} = 0.4$ | **78.27** | 72.35 | 61.23 | 52.59 | 72.59 |
| Image | $\rho_{+1} = \rho_{-1} = 0.2$ | **79.55** | **79.55** | 70.66 | 77.76 | **80.51** |
| | $\rho_{+1} = 0.3, \rho_{-1} = 0.1$ | 74.05 | 74.05 | 68.94 | **79.39** | **81.02** |
| | $\rho_{+1} = \rho_{-1} = 0.4$ | **73.73** | **73.73** | 63.66 | 69.61 | 70.79 |
| Spambase | $\rho_{+1} = \rho_{-1} = 0.2$ | 87.03 | 87.03 | 40.04 | **88.67** | 62.22 |
| | $\rho_{+1} = 0.3, \rho_{-1} = 0.1$ | **88.61** | **88.61** | 39.46 | 76.80 | 63.33 |
| | $\rho_{+1} = \rho_{-1} = 0.4$ | **81.28** | **81.28** | 39.17 | **82.03** | 66.00 |

Table 2: $\mathcal{U}_{Acc}$ measure of classification (linear) algorithms on UCI benchmark data sets. Entries within 1% from the best in each row are in bold. All the methods use *linear* kernel. *All method-specific parameters are estimated through cross-validation.* We show the best performing NHERD variant ('project' and 'exact') in each case.

*Using Gaussian kernel.* For kernelized algorithms, we set the Gaussian kernel width parameter $\gamma$ to $1/d$ where $d$ is the dimensionality of data (the default parameter setting in `libsvm`). The results comparing SVM based methods are presented in Tables 4 and 7 for accuracy and AM measure respectively. We see a similar trend in performances as in the case of linear kernel. Note that the NHERD method is not kernelizable, so the results are omitted.

Overall, the experimental results support the theoretical guarantees; we observe that the proposed methods are competitive and are able to tolerate moderate to high amounts of label noise in the data.

| Data set | Noise rates | Logistic Regression | Approach 1: (3) with $\ell_{\log}$ | Approach 2: (8) with $\ell_{\log}$ |
|---|---|---|---|---|
| Breast cancer | $\rho_{+1} = \rho_{-1} = 0.2$ | 65.86 | **70.12** | 66.40 |
| | $\rho_{+1} = 0.3, \rho_{-1} = 0.1$ | 64.11 | **70.07** | **69.46** |
| | $\rho_{+1} = \rho_{-1} = 0.4$ | 59.51 | **67.79** | 56.43 |
| Diabetes | $\rho_{+1} = \rho_{-1} = 0.2$ | 73.52 | **76.04** | 73.52 |
| | $\rho_{+1} = 0.3, \rho_{-1} = 0.1$ | 71.96 | **75.52** | 72.48 |
| | $\rho_{+1} = \rho_{-1} = 0.4$ | **67.62** | 65.89 | **66.75** |
| Thyroid | $\rho_{+1} = \rho_{-1} = 0.2$ | 82.54 | **87.80** | 82.54 |
| | $\rho_{+1} = 0.3, \rho_{-1} = 0.1$ | **82.26** | 80.34 | **82.23** |
| | $\rho_{+1} = \rho_{-1} = 0.4$ | 77.28 | **83.10** | 76.36 |
| German | $\rho_{+1} = \rho_{-1} = 0.2$ | 66.33 | **71.80** | 66.33 |
| | $\rho_{+1} = 0.3, \rho_{-1} = 0.1$ | 66.93 | **71.40** | 68.33 |
| | $\rho_{+1} = \rho_{-1} = 0.4$ | 55.87 | **67.19** | 55.41 |
| Heart | $\rho_{+1} = \rho_{-1} = 0.2$ | 81.23 | **82.96** | 81.23 |
| | $\rho_{+1} = 0.3, \rho_{-1} = 0.1$ | 81.73 | **84.44** | 81.73 |
| | $\rho_{+1} = \rho_{-1} = 0.4$ | **73.58** | 57.04 | **73.58** |
| Image | $\rho_{+1} = \rho_{-1} = 0.2$ | **82.90** | **82.45** | **82.90** |
| | $\rho_{+1} = 0.3, \rho_{-1} = 0.1$ | **82.07** | **82.55** | **82.07** |
| | $\rho_{+1} = \rho_{-1} = 0.4$ | **76.25** | 63.47 | **76.25** |
| Spambase | $\rho_{+1} = \rho_{-1} = 0.2$ | 87.72 | **89.80** | 87.72 |
| | $\rho_{+1} = 0.3, \rho_{-1} = 0.1$ | 72.37 | **89.28** | 72.37 |
| | $\rho_{+1} = \rho_{-1} = 0.4$ | 79.88 | **80.22** | 79.88 |

Table 3: $\mathcal{U}_{Acc}$ measure of logistic loss based classification algorithms on UCI benchmark data sets. Entries within 1% from the best in each row are in bold. All the methods use *linear* kernel. *All method-specific parameters are estimated through cross-validation.*

| Data set | Noise rates | SVM | C-SVM | PAM | RP |
|---|---|---|---|---|---|
| Breast cancer | $\rho_{+1} = \rho_{-1} = 0.2$ | **70.77** | **70.77** | 67.45 | 65.91 |
| | $\rho_{+1} = 0.3, \rho_{-1} = 0.1$ | 71.02 | 71.02 | **73.31** | 70.01 |
| | $\rho_{+1} = \rho_{-1} = 0.4$ | 62.64 | 62.64 | 60.56 | **63.92** |
| Diabetes | $\rho_{+1} = \rho_{-1} = 0.2$ | **74.91** | 73.35 | **74.65** | 72.40 |
| | $\rho_{+1} = 0.3, \rho_{-1} = 0.1$ | **73.87** | 73.09 | 72.66 | 68.14 |
| | $\rho_{+1} = \rho_{-1} = 0.4$ | 55.30 | 52.86 | 63.45 | **65.19** |
| Thyroid | $\rho_{+1} = \rho_{-1} = 0.2$ | **92.23** | **92.23** | 91.92 | 83.53 |
| | $\rho_{+1} = 0.3, \rho_{-1} = 0.1$ | **84.09** | **84.09** | **85.33** | 80.06 |
| | $\rho_{+1} = \rho_{-1} = 0.4$ | 73.86 | 73.86 | **82.56** | 84.43 |
| German | $\rho_{+1} = \rho_{-1} = 0.2$ | **74.20** | **74.20** | 73.80 | 72.14 |
| | $\rho_{+1} = 0.3, \rho_{-1} = 0.1$ | 70.40 | 70.40 | 70.67 | **72.60** |
| | $\rho_{+1} = \rho_{-1} = 0.4$ | **61.45** | **61.45** | 59.73 | 59.52 |
| Heart | $\rho_{+1} = \rho_{-1} = 0.2$ | 66.17 | 70.86 | **78.27** | **78.27** |
| | $\rho_{+1} = 0.3, \rho_{-1} = 0.1$ | **77.04** | **77.04** | 73.83 | 68.64 |
| | $\rho_{+1} = \rho_{-1} = 0.4$ | 60.00 | 60.74 | 67.41 | 68.89 |
| Image | $\rho_{+1} = \rho_{-1} = 0.2$ | **94.09** | **94.09** | 92.36 | 80.50 |
| | $\rho_{+1} = 0.3, \rho_{-1} = 0.1$ | **91.50** | **91.50** | 86.26 | 73.86 |
| | $\rho_{+1} = \rho_{-1} = 0.4$ | **81.11** | **81.11** | 80.38 | 75.29 |
| Spambase | $\rho_{+1} = \rho_{-1} = 0.2$ | **78.41** | **78.41** | 77.30 | 59.94 |
| | $\rho_{+1} = 0.3, \rho_{-1} = 0.1$ | **75.16** | **75.16** | **75.46** | 57.29 |
| | $\rho_{+1} = \rho_{-1} = 0.4$ | 62.72 | **65.20** | 63.55 | 55.01 |

Table 4: $\mathcal{U}_{Acc}$ measure of classification (kernelized) algorithms on UCI benchmark data sets. Entries within 1% from the best in each row are in bold. All the methods use *Gaussian* kernel with width $\gamma = 1/d$ (where $d$ is the number of dimensions). *All method-specific parameters are estimated through cross-validation.* NHERD algorithm is excluded as it is not kernelizable.

| Data set | Noise rates | SVM | C-SVM | PAM | NHERD | RP |
|---|---|---|---|---|---|---|
| Breast cancer | $\rho_{+1} = \rho_{-1} = 0.2$ | 50.82 | 50.82 | 62.94 | **66.14** | 37.58 |
| | $\rho_{+1} = 0.3, \rho_{-1} = 0.1$ | 57.48 | 57.48 | 59.94 | **64.28** | 62.69 |
| | $\rho_{+1} = \rho_{-1} = 0.4$ | 52.59 | 50.83 | **56.52** | **56.21** | **56.02** |
| Diabetes | $\rho_{+1} = \rho_{-1} = 0.2$ | 70.85 | 70.85 | 69.90 | **74.48** | 72.17 |
| | $\rho_{+1} = 0.3, \rho_{-1} = 0.1$ | **75.26** | **75.26** | 66.19 | **76.66** | 73.80 |
| | $\rho_{+1} = \rho_{-1} = 0.4$ | 63.15 | 63.15 | 60.16 | **71.88** | 69.00 |
| Thyroid | $\rho_{+1} = \rho_{-1} = 0.2$ | 72.16 | 72.16 | 67.25 | **77.67** | 74.16 |
| | $\rho_{+1} = 0.3, \rho_{-1} = 0.1$ | 79.69 | 79.69 | 55.05 | **83.99** | 74.05 |
| | $\rho_{+1} = \rho_{-1} = 0.4$ | 64.23 | 64.23 | 55.45 | **82.97** | 66.62 |
| German | $\rho_{+1} = \rho_{-1} = 0.2$ | 62.15 | 62.15 | 64.68 | **70.64** | 67.72 |
| | $\rho_{+1} = 0.3, \rho_{-1} = 0.1$ | 69.59 | 69.59 | 63.09 | **70.45** | 65.23 |
| | $\rho_{+1} = \rho_{-1} = 0.4$ | 54.13 | 53.51 | 54.62 | 54.70 | **56.00** |
| Heart | $\rho_{+1} = \rho_{-1} = 0.2$ | **81.82** | **81.82** | 75.05 | **82.97** | 75.99 |
| | $\rho_{+1} = 0.3, \rho_{-1} = 0.1$ | 77.98 | 77.98 | 73.10 | **82.19** | 78.46 |
| | $\rho_{+1} = \rho_{-1} = 0.4$ | **76.42** | 67.81 | 66.16 | 52.07 | 72.85 |
| Image | $\rho_{+1} = \rho_{-1} = 0.2$ | 76.43 | 76.43 | 67.29 | 76.75 | **79.23** |
| | $\rho_{+1} = 0.3, \rho_{-1} = 0.1$ | 75.70 | 75.70 | 67.13 | **80.21** | 76.87 |
| | $\rho_{+1} = \rho_{-1} = 0.4$ | **69.68** | **69.68** | 58.03 | **70.64** | **70.68** |
| Spambase | $\rho_{+1} = \rho_{-1} = 0.2$ | 85.88 | 85.88 | 50.00 | **88.62** | 61.19 |
| | $\rho_{+1} = 0.3, \rho_{-1} = 0.1$ | **88.27** | **88.27** | 52.07 | 80.06 | 67.67 |
| | $\rho_{+1} = \rho_{-1} = 0.4$ | 78.81 | 78.81 | 50.00 | **81.51** | 63.26 |

Table 5: $\mathcal{U}_{AM}$ measure of classification (linear) algorithms on UCI benchmark data sets. Entries within 1% from the best in each row are in bold. All the methods use *linear* kernel. *All method-specific parameters are estimated through cross-validation.* We show the best performing NHERD variant ('project' and 'exact') in each case.

| Data set | Noise rates | Logistic Regression | Approach 1: (3) with $\ell_{\log}$ | Approach 2: (8) with $\ell_{\log}$ |
|---|---|---|---|---|
| Breast cancer | $\rho_{+1} = \rho_{-1} = 0.2$ | **65.95** | 59.58 | **65.20** |
| | $\rho_{+1} = 0.3, \rho_{-1} = 0.1$ | 61.61 | 56.28 | **65.75** |
| | $\rho_{+1} = \rho_{-1} = 0.4$ | **57.11** | 51.50 | 54.50 |
| Diabetes | $\rho_{+1} = \rho_{-1} = 0.2$ | **74.70** | 63.37 | **74.70** |
| | $\rho_{+1} = 0.3, \rho_{-1} = 0.1$ | **73.38** | 63.13 | **73.74** |
| | $\rho_{+1} = \rho_{-1} = 0.4$ | **68.18** | 56.07 | **67.19** |
| Thyroid | $\rho_{+1} = \rho_{-1} = 0.2$ | 78.70 | **82.42** | 78.70 |
| | $\rho_{+1} = 0.3, \rho_{-1} = 0.1$ | **79.28** | 68.04 | **79.38** |
| | $\rho_{+1} = \rho_{-1} = 0.4$ | **73.46** | 53.19 | **72.41** |
| German | $\rho_{+1} = \rho_{-1} = 0.2$ | **69.24** | 67.47 | **69.24** |
| | $\rho_{+1} = 0.3, \rho_{-1} = 0.1$ | **69.59** | 53.87 | **70.44** |
| | $\rho_{+1} = \rho_{-1} = 0.4$ | **57.07** | 51.50 | **56.65** |
| Heart | $\rho_{+1} = \rho_{-1} = 0.2$ | **81.48** | **80.92** | **81.48** |
| | $\rho_{+1} = 0.3, \rho_{-1} = 0.1$ | 81.73 | **83.37** | 81.73 |
| | $\rho_{+1} = \rho_{-1} = 0.4$ | **74.13** | 51.59 | **74.13** |
| Image | $\rho_{+1} = \rho_{-1} = 0.2$ | **81.79** | 80.23 | **81.79** |
| | $\rho_{+1} = 0.3, \rho_{-1} = 0.1$ | **81.12** | **81.18** | **81.12** |
| | $\rho_{+1} = \rho_{-1} = 0.4$ | **75.87** | 56.60 | **75.87** |
| Spambase | $\rho_{+1} = \rho_{-1} = 0.2$ | **88.38** | **89.05** | **88.38** |
| | $\rho_{+1} = 0.3, \rho_{-1} = 0.1$ | 76.78 | **88.27** | 76.78 |
| | $\rho_{+1} = \rho_{-1} = 0.4$ | **80.97** | 75.80 | **80.97** |

Table 6: $\mathcal{U}_{AM}$ measure of logistic loss based classification algorithms on UCI benchmark data sets. Entries within 1% from the best in each row are in bold. All the methods use *linear* kernel. *All method-specific parameters are estimated through cross-validation.*

| Data set | Noise rates | SVM | C-SVM | PAM | RP |
|---|---|---|---|---|---|
| Breast cancer | $\rho_{+1} = \rho_{-1} = 0.2$ | 56.28 | 56.28 | **58.02** | 54.83 |
| | $\rho_{+1} = 0.3, \rho_{-1} = 0.1$ | 56.96 | 56.96 | 57.08 | **58.15** |
| | $\rho_{+1} = \rho_{-1} = 0.4$ | 50.84 | 50.84 | 49.84 | **52.89** |
| Diabetes | $\rho_{+1} = \rho_{-1} = 0.2$ | **70.00** | 66.68 | **70.07** | 68.20 |
| | $\rho_{+1} = 0.3, \rho_{-1} = 0.1$ | **73.98** | 71.29 | **72.90** | 69.32 |
| | $\rho_{+1} = \rho_{-1} = 0.4$ | 56.75 | 52.04 | 58.48 | **61.41** |
| Thyroid | $\rho_{+1} = \rho_{-1} = 0.2$ | **88.87** | **88.87** | **89.77** | 84.72 |
| | $\rho_{+1} = 0.3, \rho_{-1} = 0.1$ | 74.97 | 74.97 | **78.26** | 68.13 |
| | $\rho_{+1} = \rho_{-1} = 0.4$ | 66.70 | 66.70 | 74.53 | **80.59** |
| German | $\rho_{+1} = \rho_{-1} = 0.2$ | **63.51** | **63.51** | **64.46** | **64.41** |
| | $\rho_{+1} = 0.3, \rho_{-1} = 0.1$ | **67.79** | **67.79** | **67.71** | 65.86 |
| | $\rho_{+1} = \rho_{-1} = 0.4$ | 52.60 | 52.60 | 53.27 | **54.61** |
| Heart | $\rho_{+1} = \rho_{-1} = 0.2$ | 63.98 | 68.86 | **77.41** | **77.65** |
| | $\rho_{+1} = 0.3, \rho_{-1} = 0.1$ | **77.61** | **77.61** | 74.99 | 69.59 |
| | $\rho_{+1} = \rho_{-1} = 0.4$ | 57.80 | 56.44 | **66.71** | 65.85 |
| Image | $\rho_{+1} = \rho_{-1} = 0.2$ | **93.51** | **93.51** | 91.45 | 80.47 |
| | $\rho_{+1} = 0.3, \rho_{-1} = 0.1$ | **92.00** | **92.00** | 87.40 | 75.42 |
| | $\rho_{+1} = \rho_{-1} = 0.4$ | **78.90** | **78.90** | 77.78 | 75.58 |
| Spambase | $\rho_{+1} = \rho_{-1} = 0.2$ | **77.11** | **77.11** | 75.82 | 56.44 |
| | $\rho_{+1} = 0.3, \rho_{-1} = 0.1$ | **77.47** | **77.47** | **77.75** | 57.42 |
| | $\rho_{+1} = \rho_{-1} = 0.4$ | 55.52 | **59.57** | **60.84** | 53.27 |

Table 7: $\mathcal{U}_{AM}$ measure of classification (kernelized) algorithms on UCI benchmark data sets. Entries within 1% from the best in each row are in bold. All the methods use *Gaussian* kernel with width $\gamma = 1/d$ (where $d$ is the number of dimensions). *All method-specific parameters are estimated through cross-validation.* NHERD algorithm is excluded as it is not kernelizable.

### 6.3 Knowledge of noise rates

The proposed algorithms require the knowledge of noise rates $\rho_{+1}$ and $\rho_{-1}$. However, in practice, we do not know the true value of noise rates, and therefore we resort to cross-validating the values in our experiments. In some cases (and domains), we may be able to approximately specify noise rates. This motivates our study presented in Figure 3. True noise rates $\rho_{+1} = \rho_{-1} = \rho$ are misspecified as $(\rho_{+1} \pm \epsilon, \rho_{-1} \pm \epsilon)$ for $\epsilon \in \{0.1, 0.2, 0.3, 0.4\}$. The ratio between the average accuracy for a given $\epsilon$ and the accuracy at $\epsilon = 0$, i.e. when true noise rates are specified, is a measure of sensitivity of the algorithms to $\epsilon$-misspecification of noise rates. We would want the ratio to be close to 1 for a given $\epsilon$, which would suggest that the method is fairly robust with respect to the $\epsilon$-misspecification. The results in Figure 3 show that the proposed methods are robust to $\epsilon$-misspecification of noise rates, which in turn suggests that our methods can find better use in applications where labels can be noisy *and* noise rates are approximately known, without resorting to ad-hoc cross-validation procedures on the noisy data. We emphasize here that in case the true noise rates are known, our methods can benefit from that knowledge as observed from our experiments, whereas the competitive methods *cannot* as they do not involve noise rates.

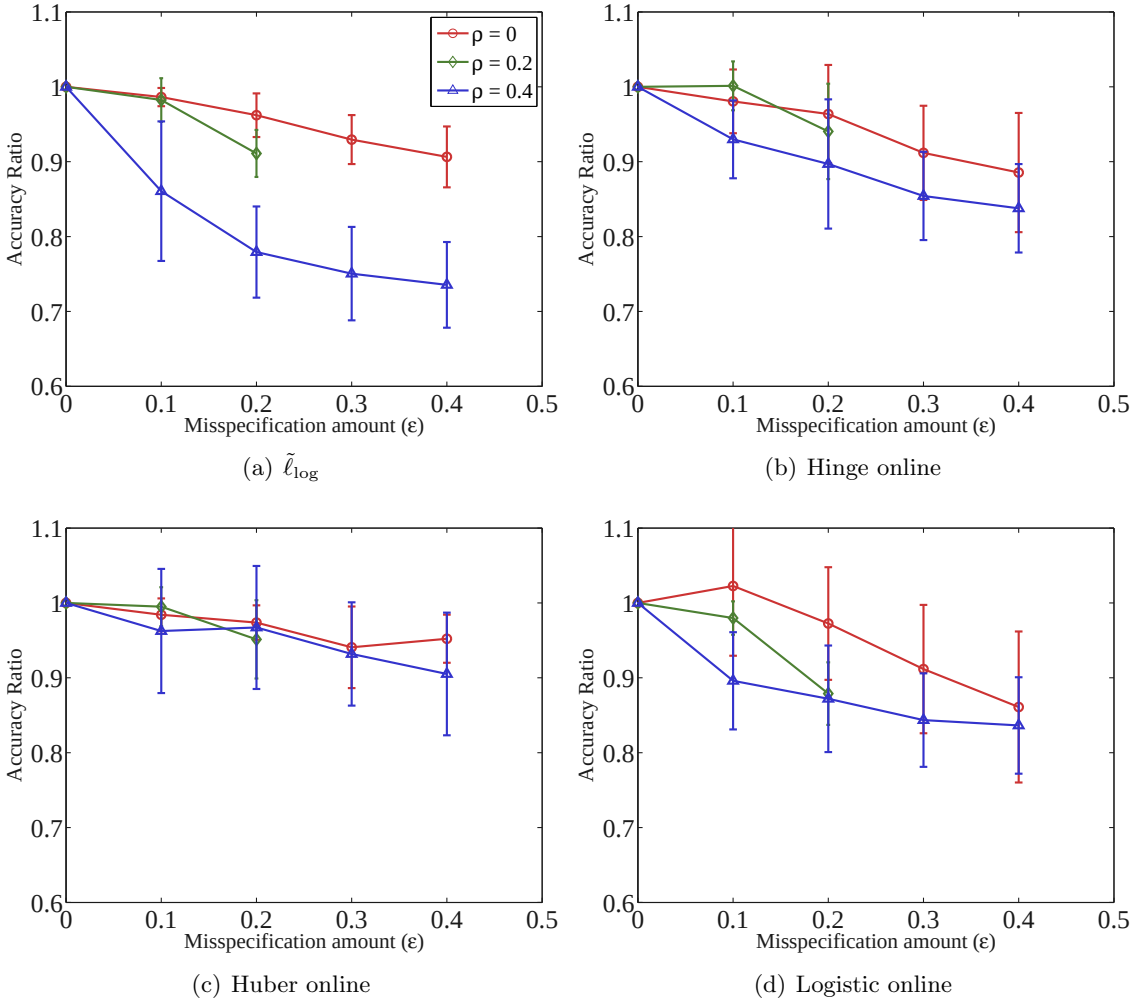
## 7. Conclusions and Future Work

We addressed learning in the presence of asymmetric random label noise with respect to general cost-sensitive utilities. We have obtained general theoretical results as well as efficient algorithms for this setting using the methods of unbiased estimators and weighted loss functions. The proposed algorithms are easy to implement and the classification performance is encouraging even at high noise rates and in particular is competitive with state-of-the-art methods on benchmark data. Our developments provide a new family of methods that can be applied to the positive-unlabeled learning problem (Elkan and Noto, 2008), but the implications of our methods for this setting should be carefully analyzed. We could consider harder noise models such as label noise depending on the example, and nastier variants of label noise where labels to flip are chosen adversarially.

Our analysis in this paper covers cost-sensitive classification losses, but there are other measures used in practice such as $F_\beta$, that are not covered by our family. Consistent learning for such general performance measures is beginning to be understood in the noise-free setting (Koyejo et al., 2014; Narasimhan et al., 2014). It would be interesting to see if we can extend some of the ideas in this paper to more general utility measures. It will also be of interest to extend our methods to deal with label noise in more general learning problems such as classification with a reject option, multiclass classification, learning with partial labels, learning to rank, and multilabel classification. Some of these extensions have begun to occur (van Rooyen and Williamson, 2015) but others are yet to be explored.

## Acknowledgments

We gratefully acknowledge the support of NSF under grants CCF-1320746 and CCF-1117055; P.R. acknowledges the support of ARO via W911NF-12-1-0390 and NSF via IIS-1149803, IIS-1320894. A.T. acknowledges the support of NSF via CCF-1422157.




(a) $\tilde{\ell}_{\log}$ (b) Hinge online

(c) Huber online (d) Logistic online

Figure 3: Study of sensitivity of batch ($\tilde{\ell}_{\log}$) and online (Hinge, Huber and Logistic) methods (Algorithm 1) to specification of noise rates $\rho_{+1}$ and $\rho_{-1}$. True noise rates $\rho_{+1} = \rho_{-1} = \rho$ are misspecified as $(\rho_{+1} \pm \epsilon, \rho_{-1} \pm \epsilon)$ for $\epsilon \in \{0.1, 0.2, 0.3, 0.4\}$. The ratio between the average accuracy for a given $\epsilon$ and the accuracy at $\epsilon = 0$, i.e. when true noise rates are specified, is plotted for different values of noise rates $\rho$. The ratio is computed for each of the 6 UCI data sets in Table 1 and the mean and the standard deviation of the ratios are shown. Ratio being equal to 1 for a given $\epsilon$ means that the performance of the algorithm, on average, is unaltered by misspecification of noise rates up to $\epsilon$. As expected, the ratio decreases, i.e. the algorithms perform worse as $\epsilon$ increases. Most of the ratios being close to 1 suggests that the proposed methods are fairly robust with respect to $\epsilon$-misspecification of noise rates.

## References

D. Angluin and P. Laird. Learning from noisy examples. *Mach. Learn.*, 2(4):343–370, 1988.

Javed A. Aslam and Scott E. Decatur. On the sample complexity of noise-tolerant learning. *Inf. Process. Lett.*, 57(4):189–195, 1996.

Peter L. Bartlett, Michael I. Jordan, and Jon D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

Shai Ben-David, Dávid Pál, and Shai Shalev-Shwartz. Agnostic online learning. In *Conference on Learning Theory (COLT)*, 2009.

Battista Biggio, Blaine Nelson, and Pavel Laskov. Support vector machines under adversarial label noise. *J. Mach. Learn. Res. (JMLR)*, 20:97–112, 2011.

Gilles Blanchard and Clayton Scott. Decontamination of mutually contaminated models. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1–9, 2014.

A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *Computational Learning Theory (COLT)*, pages 92–100. ACM, 1998.

Tom Bylander. Learning linear threshold functions in the presence of classification noise. In *Conference on Learning Theory (COLT)*, pages 340–347, NY, USA, 1994. ACM.

Tom Bylander. Learning noisy linear threshold functions. *Technical Report*, 1998.

Nicolò Cesa-Bianchi, Eli Dichterman, Paul Fischer, Eli Shamir, and Hans Ulrich Simon. Sample-efficient strategies for learning in the presence of noise. *J. ACM*, 46(5):684–719, 1999.

Nicolò Cesa-Bianchi, Shai Shalev-Shwartz, and Ohad Shamir. Online learning of noisy data. *IEEE Transactions on Information Theory*, 57(12):7907–7931, 2011.

Edith Cohen. Learning noisy perceptrons by a perceptron in polynomial time. In *Foundations of Computer Science*, pages 514–523. IEEE, 1997.

K. Crammer and D. Lee. Learning via Gaussian Herding. In *Neural Information Processing Systems (NIPS)*, pages 451–459, 2010.

Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer. Online passive-aggressive algorithms. *J. Mach. Learn. Res. (JMLR)*, 7:551–585, 2006.

Koby Crammer, Alex Kulesza, and Mark Dredze. Adaptive regularization of weight vectors. In *Neural Information Processing Systems (NIPS)*, pages 414–422, 2009.

Mark Dredze, Koby Crammer, and Fernando Pereira. Confidence-weighted linear classification. In *International Conference on Machine Learning (ICML)*, pages 264–271, 2008.

C. Elkan and K. Noto. Learning classifiers from only positive and unlabeled data. In *Intl. Conf. on Knowledge Discovery and Data Mining (KDD)*, pages 213–220, 2008.

Charles Elkan. The foundations of cost-sensitive learning. In *International Joint Conference on Artificial Intelligence (IJCAI)*, IJCAI'01, pages 973–978, 2001.

Yoav Freund. A more robust boosting algorithm, 2009. preprint `arXiv:0905.2138 [stat.ML]` available at `http://arxiv.org/abs/0905.2138`.

Aritra Ghosh, Naresh Manwani, and P. S. Sastry. Making risk minimization tolerant to label noise. *CoRR*, abs/1403.3610, 2014. URL `http://arxiv.org/abs/1403.3610`.

T. Graepel and R. Herbrich. The kernel Gibbs sampler. In *Neural Information Processing Systems (NIPS)*, pages 514–520, 2000.

Roni Khardon and Gabriel Wachman. Noise tolerant variants of the perceptron algorithm. *J. Mach. Learn. Res. (JMLR)*, 8:227–248, 2007.

Oluwasanmi O Koyejo, Nagarajan Natarajan, Pradeep K Ravikumar, and Inderjit S Dhillon. Consistent binary classification with generalized performance metrics. In *Neural Information Processing Systems (NIPS)*, pages 2744–2752, 2014.

Neil D. Lawrence and Bernhard Schölkopf. Estimating a kernel Fisher discriminant in the presence of label noise. In *International Conference on Machine Learning (ICML)*, pages 306–313, 2001.

Bing Liu, Yang Dai, Xiaoli Li, Wee Sun Lee, and Philip S Yu. Building text classifiers using positive and unlabeled examples. In *ICDM 2003.*, pages 179–186. IEEE, 2003.

Philip M. Long and Rocco A. Servedio. Random classification noise defeats all convex potential boosters. *Mach. Learn.*, 78(3):287–304, 2010.

Yves Lucet. What shape is your conjugate? a survey of computational convex analysis and its applications. *SIAM Rev.*, 52(3):505–542, August 2010. ISSN 0036-1445.

Naresh Manwani and P. S. Sastry. Noise tolerance under risk minimization. *IEEE Trans. Syst. Man and Cybern. Part B*, 2013. URL: `http://arxiv.org/abs/1109.5231`.

Aditya Menon, Harikrishna Narasimhan, Shivani Agarwal, and Sanjay Chawla. On the statistical consistency of algorithms for binary classification under class imbalance. In *International Conference on Machine Learning (ICML)*, pages 603–611, 2013.

Fantine Mordelet and J-P Vert. A bagging SVM to learn from positive and unlabeled examples. *Pattern Recognition Letters*, 37:201–209, 2014.

Harikrishna Narasimhan, Rohit Vaish, and Shivani Agarwal. On the statistical consistency of plug-in classifiers for non-decomposable performance measures. In *Neural Information Processing Systems (NIPS)*, pages 1493–1501, 2014.

Nagarajan Natarajan, Inderjit Dhillon, Pradeep Ravikumar, and Ambuj Tewari. Learning with noisy labels. In *Neural Information Processing Systems (NIPS)*, pages 1196–1204, 2013.

A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM J. on Opt.*, 19(4):1574–1609, 2009.

David F. Nettleton, A. Orriols-Puig, and A. Fornells. A study of the effect of different types of noise on the precision of supervised learning techniques. *Artif. Intell. Rev.*, 33 (4):275–306, 2010.

C. Scott. A rate of convergence for mixture proportion estimation, with application to learning from noisy labels. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 38 of *JMLR Workshop and Conference Proceedings*, pages 838–846, 2015.

Clayton Scott. Calibrated asymmetric surrogate losses. *Electronic J. of Stat.*, 6:958–992, 2012.

Clayton Scott, Gilles Blanchard, and Gregory Handy. Classification with asymmetric label noise: Consistency and maximal denoising. *Conference on Learning Theory (COLT)*, 30: 489–511, 2013.

G. Stempfel and L. Ralaivola. Learning kernel perceptrons on noisy data using random projections. In *Algorithmic Learning Theory (ALT)*, pages 328–342. Springer, 2007.

G. Stempfel, L. Ralaivola, and F. Denis. Learning from noisy data using hyperplane sampling and sample averages. 2007.

Guillaume Stempfel and Liva Ralaivola. Learning SVMs from sloppily labeled data. In *Artificial Neural Networks*, pages 884–893. Springer-Verlag, 2009.

Brendan van Rooyen and Robert C Williamson. Learning in the presence of corruption, 2015. arXiv preprint arXiv:1504.00091.

Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *International Conference on Machine Learning (ICML)*, pages 928–936, 2003.