Learning with Noisy Labels

Nagarajan Natarajan **Inderjit S. Dhillon** **Pradeep Ravikumar**
Department of Computer Science, University of Texas, Austin.
{naga86,inderjit,pradeepr}@cs.utexas.edu

Ambuj Tewari
Department of Statistics, University of Michigan, Ann Arbor.
tewaria@umich.edu

Abstract

In this paper, we theoretically study the problem of binary classification in the presence of random classification noise — the learner, instead of seeing the true labels, sees labels that have independently been flipped with some small probability. Moreover, random label noise is *class-conditional* — the flip probability depends on the class. We provide two approaches to suitably modify any given surrogate loss function. First, we provide a simple unbiased estimator of any loss, and obtain performance bounds for empirical risk minimization in the presence of iid data with noisy labels. If the loss function satisfies a simple symmetry condition, we show that the method leads to an efficient algorithm for empirical minimization. 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 empirical risk bounds. This approach has a very remarkable consequence — methods used in practice such as biased SVM and weighted logistic regression are provably noise-tolerant. On a synthetic non-separable dataset, our methods achieve over 88% accuracy even when 40% of the labels are corrupted, and are competitive with respect to recently proposed methods for dealing with label noise in several benchmark datasets.

1 Introduction

Designing supervised learning algorithms that can learn from data sets with noisy labels is a problem of great practical importance. Here, by noisy labels, we refer to the setting where an adversary has deliberately corrupted the labels [Biggio et al., 2011], which otherwise arise from some "clean" distribution; learning from only positive and unlabeled data [Elkan and Noto, 2008] can also be cast in this setting. Given the importance of learning from such noisy labels, a great deal of practical work has been done on the problem (see, for instance, the survey article by Nettleton et al. [2010]). The theoretical machine learning community has also investigated the problem of learning from noisy labels. 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 Littestone 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. Learning with convex losses has been addressed only under limiting assumptions like separability or uniform noise rates [Manwani and Sastry, 2013].

In this paper, we consider risk minimization in the presence of *class-conditional* random label noise (abbreviated CCN). The data consists of iid samples from an underlying "clean" distribution D. The learning algorithm sees samples drawn from a noisy version D_ρ of D — 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. To this end, we develop two methods for suitably modifying *any given surrogate loss function* ℓ, and show that minimizing the sample average of the modified proxy loss function

$\tilde{\ell}$ leads to provable risk bounds where the risk is calculated using the original loss ℓ on the clean distribution.

In our first approach, the modified or proxy loss is an unbiased estimate of the loss function. The idea of using unbiased estimators is well-known in stochastic optimization [Nemirovski et al., 2009], and regret bounds can be obtained for learning with noisy labels in an online learning setting (See Appendix B). Nonetheless, we bring out some important aspects of using unbiased estimators of loss functions for empirical risk minimization 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 6).

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 very remarkable result that appropriately weighted losses like biased SVMs studied by Liu et al. [2003] are robust to CCN.

The main results and the contributions of the paper are summarized below:

1. To the best of our knowledge, we are the first to provide guarantees for risk minimization under random label noise in the general setting of convex surrogates, without any assumptions on the true distribution.
2. We provide two different approaches to suitably modifying any given surrogate loss function, that surprisingly lead to very similar risk bounds (Theorems 3 and 11). These general results include some existing results for random classification noise as special cases.
3. We resolve an elusive theoretical gap in the understanding of practical methods like biased SVM and weighted logistic regression — they are provably noise-tolerant (Theorem 11).
4. Our proxy losses are easy to compute — both the methods yield efficient algorithms.
5. Experiments on benchmark datasets show that the methods are robust even at high noise rates.

The outline of the paper is as follows. We introduce the problem setting and terminology in Section 2. In Section 3, we give our first main result concerning the method of unbiased estimators. In Section 4, we give our second and third main results for certain weighted loss functions. We present experimental results on synthetic and benchmark data sets in Section 5.

1.1 Related Work

Starting from the work of Bylander [1994], many noise tolerant versions of the perceptron algorithm have been developed. 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. A Bayesian approach to the problem of noisy labels is taken by Graepel and Herbrich [2000] and Lawrence and Schölkopf [2001]. As Adaboost is very sensitive to label noise, random label noise has also been considered in the context of boosting. Long and Servedio [2010] prove that any method based on a convex potential is inherently ill-suited to random label noise. Freund [2009] proposes a boosting algorithm based on a non-convex potential that is empirically seen to be robust against random label noise.

Stempfel and Ralaivola [2009] proposed the minimization of an unbiased proxy for the case of the hinge loss. However the hinge loss leads to a non-convex problem. Therefore, they proposed heuristic minimization approaches for which no theoretical guarantees were provided (We address the issue in Section 3.1). Cesa-Bianchi et al. [2011] focus on the online learning algorithms where they only need unbiased estimates of the gradient of the loss 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. In particular, their techniques break down for non-smooth losses such as the hinge loss. In contrast, we show that unbiased estimation is always possible in the more benign random classification noise setting. 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 it holds promise 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, Scott et al. [2013] study the problem of classification under 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. Furthermore, they do not give any efficient algorithm for the problem.

2 Problem Setup and Background

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. 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:

$$P(\tilde{Y} = -1|Y = +1) = \rho_{+1}, P(\tilde{Y} = +1|Y = -1) = \rho_{-1}, \text{ and } \rho_{+1} + \rho_{-1} < 1$$

The corrupted samples are what the learning algorithm sees. We will assume that the noise rates ρ_{+1} and ρ_{-1} are known[1] to the learner. Let the distribution of $(X, \tilde{Y})$ be D_ρ. Instances are denoted by $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^d$. Noisy labels are denoted by $\tilde{y}$.

Let $f : \mathcal{X} \to \mathbb{R}$ be some real-valued decision function. The *risk* of f w.r.t. the 0-1 loss is given by $R_D(f) = \mathbb{E}_{(X,Y)\sim D}\left[1_{\{\text{sign}(f(X))\neq Y\}}\right]$. The optimal decision function (called Bayes optimal) that minimizes R_D over all real-valued decision functions is given by $f^\star(x) = \text{sign}(\eta(x) - 1/2)$ where $\eta(x) = P(Y = 1|x)$. We denote by R^* the corresponding *Bayes risk* under the clean distribution D, i.e. $R^* = R_D(f^*)$. Let $\ell(t, y)$ denote a loss function where $t \in \mathbb{R}$ is a real-valued prediction and $y \in \{\pm 1\}$ is a label. Let $\tilde{\ell}(t, \tilde{y})$ denote a suitably modified ℓ for use with noisy labels (obtained using methods in Sections 3 and 4). It is helpful to summarize the three important quantities associated with a decision function f:

1. Empirical $\tilde{\ell}$-risk on the observed sample: $\widehat{R}_{\tilde{\ell}}(f) := \frac{1}{n}\sum_{i=1}^n \tilde{\ell}(f(X_i), \tilde{Y}_i)$.
2. As n grows, we expect $\widehat{R}_{\tilde{\ell}}(f)$ to be close to the $\tilde{\ell}$-risk under the noisy distribution D_ρ:
$$R_{\tilde{\ell},D_\rho}(f) := \mathbb{E}_{(X,\tilde{Y})\sim D_\rho}\left[\tilde{\ell}(f(X), \tilde{Y})\right] .$$
3. ℓ-risk under the "clean" distribution D: $R_{\ell,D}(f) := \mathbb{E}_{(X,Y)\sim D}\left[\ell(f(X), Y)\right]$.

Typically, ℓ is a convex function that is *calibrated* with respect to an underlying loss function such as the 0-1 loss. ℓ is said to be *classification-calibrated* [Bartlett et al., 2006] if and only if there exists a convex, invertible, nondecreasing transformation ψ_ℓ (with $\psi_\ell(0) = 0$) such that $\psi_\ell(R_D(f) - R^*) \leq R_{\ell,D}(f) - \min_f R_{\ell,D}(f)$. The interpretation is that we can control the excess 0-1 risk by controlling the excess ℓ-risk.

If f is not quantified in a minimization, then it is implicit that the minimization is over all measurable functions. 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. Proofs are provided in the Appendix A.

3 Method of Unbiased Estimators

Let $\mathcal{F} : \mathcal{X} \to \mathbb{R}$ be a fixed class of real-valued decision functions, over which the empirical risk is minimized. The method of unbiased estimators uses the noise rates to construct an unbiased estimator $\tilde{\ell}(t, \tilde{y})$ for the loss $\ell(t, y)$. However, in the experiments we will tune the noise rate parameters through cross-validation. The following key lemma tells us how to construct unbiased estimators of the loss from noisy labels.

Lemma 1. *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, $\mathbb{E}_{\tilde{y}}\left[\tilde{\ell}(t, \tilde{y})\right] = \ell(t, y)$.

[1] This is not necessary in practice. See Section 5.

We can try to learn a good predictor in the presence of label noise by minimizing the sample average

$$\hat{f} \leftarrow \operatorname*{argmin}_{f \in \mathcal{F}} \widehat{R}_{\tilde{\ell}}(f) \, .$$

By unbiasedness of $\tilde{\ell}$ (Lemma 1), we know that, for any fixed $f \in \mathcal{F}$, the above sample average converges to $R_{\ell,D}(f)$ 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 $\tilde{\ell}$. The price to pay for this is L_ρ, the Lipschitz constant of $\tilde{\ell}$.

Lemma 2. *Let $\ell(t, y)$ be L-Lipschitz in t (for every y). Then, with probability at least $1 - \delta$,*

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

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 $L_\rho \leq 2L/(1 - \rho_{+1} - \rho_{-1})$ is the Lipschitz constant of $\tilde{\ell}$. Note that ϵ_i's are iid Rademacher (symmetric Bernoulli) random variables.

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.

Theorem 3 (Main Result 1)**.** *With probability at least $1 - \delta$,*

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

Furthermore, if ℓ is classification-calibrated*, there exists a nondecreasing function ζ_ℓ with $\zeta_\ell(0) = 0$ such that,*

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

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 misclassification probability of $\hat{f}$ approaches the Bayes risk of 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 $\tilde{\ell}$ is convex. Next, we address the issue of convexity of $\tilde{\ell}$.

3.1 Convex losses and their estimators

Note that the loss $\tilde{\ell}$ may not be convex even if we start with a convex ℓ. 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 4. *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.

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 ℓ 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), we could use a stochastic mirror descent type algorithm [Nemirovski et al., 2009] to arrive at risk bounds (See Appendix B) similar to Theorem 3. Then, we only need the expected loss to be convex and therefore

ℓ_{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 2 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 5. *Let $F : \mathcal{W} \to \mathbb{R}$ be a non-convex function defined on function class $\mathcal{W}$ such it is ε-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 2ε-approximate (global) minimizer of F.

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 6. *Let ℓ 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 $F(\mathbf{w}) = \widehat{R}_{\tilde{\ell}}(\mathbf{w})$. Then, with probability at least $1-\delta$, $\hat{\mathbf{w}}_{\text{approx}}$ is a 2ε-minimizer of $\widehat{R}_{\tilde{\ell}}(\cdot)$ where

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

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

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

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 Method of label-dependent costs

We develop the method of label-dependent costs from two key observations. First, the Bayes classifier for noisy distribution, denoted $\tilde{f}^*$, for the case $\rho_{+1} \neq \rho_{-1}$, simply uses a threshold different from $1/2$. Second, $\tilde{f}^*$ is the minimizer of a "label-dependent 0-1 loss" on 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. The first observation is formalized in the lemma below.

Lemma 7. *Denote $P(Y=1|X)$ by $\eta(X)$ and $P(\tilde{Y}=1|X)$ by $\tilde{\eta}(X)$. The Bayes classifier under the noisy distribution, $\tilde{f}^* = \operatorname{argmin}_f E_{(X,\tilde{Y})\sim D_\rho}\left[1_{\{\text{sign}(f(X))\neq\tilde{Y}\}}\right]$ is given by,*

$$\tilde{f}^*(x) = \text{sign}(\tilde{\eta}(x) - 1/2) = \text{sign}\left(\eta(x) - \frac{1/2 - \rho_{-1}}{1 - \rho_{+1} - \rho_{-1}}\right).$$

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:

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

We realize that the classical 0-1 loss is *unweighted*. Now, we could consider an α-weighted version of the 0-1 loss as:

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

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 α-weighted 0-1 loss, the minimizer thresholds $\eta(x)$ at α.

Lemma 8 (α-weighted Bayes optimal [Scott, 2012]). *Define U_α-risk under distribution D as*

$$R_{\alpha,D}(f) = E_{(X,Y)\sim D}[U_\alpha(f(X), Y)].$$

Then, $f^*_\alpha(x) = \text{sign}(\eta(x) - \alpha)$ minimizes U_α-risk.

Now consider the risk of f w.r.t. the α-weighted 0-1 loss under noisy distribution D_ρ:

$$R_{\alpha,D_\rho}(f) = E_{(X,\tilde{Y})\sim D_\rho}\big[U_\alpha(f(X), \tilde{Y})\big].$$

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

Theorem 9 (Main Result 2). *For the choices,*

$$\alpha^* = \frac{1 - \rho_{+1} + \rho_{-1}}{2} \text{ and } A_\rho = \frac{1 - \rho_{+1} - \rho_{-1}}{2},$$

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

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

Corollary 10. *The $\alpha^\star$-weighted Bayes optimal classifier under noisy distribution coincides with that of 0-1 loss under clean distribution:*

$$\operatorname*{argmin}_f R_{\alpha^*,D_\rho}(f) = \operatorname*{argmin}_f R_D(f) = \text{sign}(\eta(x) - 1/2).$$

4.1 Proposed Proxy Surrogate Losses

Consider any surrogate loss function ℓ; and the following decomposition:

$$\ell(t, y) = 1_{\{y=1\}}\ell_1(t) + 1_{\{y=-1\}}\ell_{-1}(t)$$

where ℓ_1 and ℓ_{-1} are partial losses of ℓ. Analogous to the 0-1 loss case, we can define α-weighted loss function (Eqn. (1)) and the corresponding α-weighted ℓ-risk. Can we hope to minimize an α-weighted ℓ-risk with respect to noisy distribution D_ρ and yet bound the excess 0-1 risk with respect to the clean distribution D? Indeed, the $\alpha^\star$ specified in Theorem 9 is precisely what we need. We are ready to state our third main result, which relies on a generalized notion of classification calibration for α-weighted losses [Scott, 2012]:

Theorem 11 (Main Result 3). *Consider the empirical risk minimization problem with noisy labels:*

$$\hat{f}_\alpha = \operatorname*{argmin}_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \ell_\alpha(f(X_i), \tilde{Y}_i).$$

Define ℓ_α as an α-weighted margin loss function of the form:

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

where $\ell : \mathbb{R} \to [0, \infty)$ is a convex loss function with Lipschitz constant L such that it is classification-calibrated (i.e. $\ell^{'}(0) < 0$). Then, for the choices α^* and A_ρ in Theorem 9, there exists a non-decreasing function $\zeta_{\ell_{\alpha^\star}}$ with $\zeta_{\ell_{\alpha^\star}}(0) = 0$, such that the following bound holds with probability at least $1 - \delta$:

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

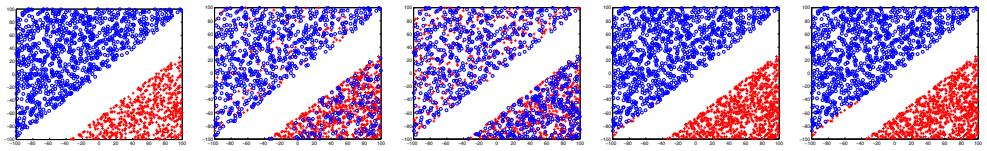
Aside from bounding excess 0-1 risk under the clean distribution, the importance of the above theorem lies in the fact that it prescribes an efficient algorithm for empirical minimization with noisy labels: ℓ_α is convex if ℓ is convex. Thus for any surrogate loss function including ℓ_{hin}, $\hat{f}_{\alpha^*}$ can be efficiently computed using the method of label-dependent costs. Note that the choice of α^* above is quite intuitive. For instance, when $\rho_{-1} \ll \rho_{+1}$ (this occurs in settings such as Liu et al. [2003] where there are only positive and unlabeled examples), $\alpha^* < 1 - \alpha^*$ and therefore mistakes on positives are penalized more than those on negatives. This makes intuitive sense since an observed negative may well have been a positive but the other way around is unlikely. In practice we do not need to know α^*, i.e. the noise rates ρ_{+1} and ρ_{-1}. The optimization problem involves just one parameter that can be tuned by cross-validation (See Section 5).

5 Experiments

We show the robustness of the proposed algorithms to increasing rates of label noise on synthetic and real-world datasets. We compare the performance of the two proposed methods with state-of-the-art methods for dealing with random classification noise. We divide each dataset (randomly) into 3 training and test sets. We use a cross-validation set to tune the parameters specific to the algorithms. Accuracy of a classification algorithm is defined as the fraction of examples in the test set classified correctly *with respect to the clean distribution*. For given noise rates ρ_{+1} and ρ_{-1}, labels of the training data are flipped accordingly and average accuracy over 3 train-test splits is computed[2]. For evaluation, we choose a representative algorithm based on each of the two proposed methods — $\tilde{\ell}_{\log}$ for the method of unbiased estimators and the widely-used C-SVM [Liu et al., 2003] method (which applies different costs on positives and negatives) for the method of label-dependent costs.

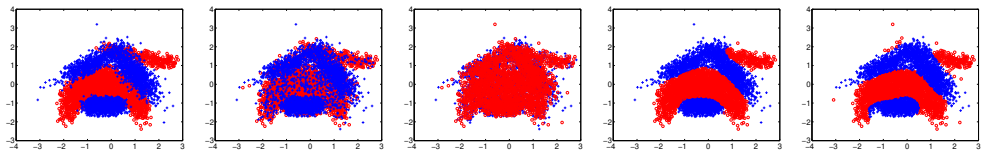
5.1 Synthetic data

First, we use the synthetic 2D linearly separable dataset 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 dataset for different noise rates. Next, we use a 2D UCI benchmark non-separable dataset ('banana'). The dataset and classification results using C-SVM (in fact, for uniform noise rates, $\alpha^* = 1/2$, so it is just the regular SVM) are shown in Figure 2. The results for higher noise rates are impressive as observed from Figures 2(d) and 2(e). The 'banana' dataset 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 $\rho_{+1} = \rho_{-1} = 0.4$.



(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).



(a) (b) (c) (d) (e)

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%.

5.2 Comparison with state-of-the-art methods on UCI benchmark

We 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

[2]Note that training and cross-validation are done on the noisy training data in our setting. To account for randomness in the flips to simulate a given noise rate, we repeat each experiment 3 times — independent corruptions of the data set for same setting of ρ_{+1} and ρ_{-1}, and present the mean accuracy over the trials.

DATASET (d, n_+, n_-)	Noise rates	$\tilde{\ell}_{\log}$	C-SVM	PAM	NHERD	RP
Breast cancer $(9, 77, 186)$	$\rho_{+1} = \rho_{-1} = 0.2$	**70.12**	67.85	**69.34**	64.90	**69.38**
	$\rho_{+1} = 0.3, \rho_{-1} = 0.1$	**70.07**	67.81	67.79	65.68	66.28
	$\rho_{+1} = \rho_{-1} = 0.4$	**67.79**	**67.79**	**67.05**	56.50	54.19
Diabetes $(8, 268, 500)$	$\rho_{+1} = \rho_{-1} = 0.2$	**76.04**	66.41	69.53	73.18	**75.00**
	$\rho_{+1} = 0.3, \rho_{-1} = 0.1$	**75.52**	66.41	65.89	**74.74**	67.71
	$\rho_{+1} = \rho_{-1} = 0.4$	65.89	65.89	65.36	**71.09**	62.76
Thyroid $(5, 65, 150)$	$\rho_{+1} = \rho_{-1} = 0.2$	87.80	94.31	**96.22**	78.49	84.02
	$\rho_{+1} = 0.3, \rho_{-1} = 0.1$	80.34	**92.46**	86.85	87.78	83.12
	$\rho_{+1} = \rho_{-1} = 0.4$	83.10	66.32	70.98	**85.95**	57.96
German $(20, 300, 700)$	$\rho_{+1} = \rho_{-1} = 0.2$	**71.80**	68.40	63.80	67.80	62.80
	$\rho_{+1} = 0.3, \rho_{-1} = 0.1$	**71.40**	68.40	67.80	67.80	67.40
	$\rho_{+1} = \rho_{-1} = 0.4$	67.19	**68.40**	67.80	54.80	59.79
Heart $(13, 120, 150)$	$\rho_{+1} = \rho_{-1} = 0.2$	**82.96**	61.48	69.63	**82.96**	72.84
	$\rho_{+1} = 0.3, \rho_{-1} = 0.1$	**84.44**	57.04	62.22	81.48	79.26
	$\rho_{+1} = \rho_{-1} = 0.4$	57.04	54.81	53.33	52.59	**68.15**
Image $(18, 1188, 898)$	$\rho_{+1} = \rho_{-1} = 0.2$	82.45	**91.95**	**92.90**	77.76	65.29
	$\rho_{+1} = 0.3, \rho_{-1} = 0.1$	82.55	**89.26**	**89.55**	79.39	70.66
	$\rho_{+1} = \rho_{-1} = 0.4$	63.47	63.47	**73.15**	69.61	64.72

Table 1: Comparative study of classification algorithms on UCI benchmark datasets. Entries within 1% from the best in each row are in bold. All the methods except NHERD variants (which are not kernelizable) use Gaussian kernel with width 1. *All method-specific parameters are estimated through cross-validation*. Proposed methods ($\tilde{\ell}_{\log}$ and C-SVM) are competitive across all the datasets. We show the best performing NHERD variant ('project' and 'exact') in each case.

[Crammer and Lee, 2010]) (*project* and *exact* variants[3]), and perceptron algorithm with margin (PAM) which was shown to be robust to label noise by Khardon and Wachman [2007]. We use the standard UCI classification datasets, preprocessed and made available by Gunnar Rätsch(`http://theoval.cmp.uea.ac.uk/matlab`). For kernelized algorithms, we use Gaussian kernel with width set to the best width obtained by tuning it for a traditional SVM on the noise-free data. For $\tilde{\ell}_{\log}$, we use ρ_{+1} and ρ_{-1} that give the best accuracy in cross-validation. For C-SVM, we fix one of the weights to 1, and tune the other. Table 1 shows the performance of the methods for different settings of noise rates. C-SVM is competitive in 4 out of 6 datasets (Breast cancer, Thyroid, German and Image), while relatively poorer in the other two. On the other hand, $\tilde{\ell}_{\log}$ is competitive in all the data sets, and performs the best more often. When about 20% labels are corrupted, uniform ($\rho_{+1} = \rho_{-1} = 0.2$) and non-uniform cases ($\rho_{+1} = 0.3, \rho_{-1} = 0.1$) have similar accuracies in all the data sets, for both C-SVM and $\tilde{\ell}_{\log}$. Overall, we observe that the proposed methods are competitive and are able to tolerate moderate to high amounts of label noise in the data. Finally, in domains where noise rates are approximately known, our methods can benefit from the knowledge of noise rates. Our analysis shows that the methods are fairly robust to misspecification of noise rates (See Appendix C for results).

6 Conclusions and Future Work

We addressed the problem of risk minimization in the presence of random classification noise, and obtained general results in the setting using the methods of unbiased estimators and weighted loss functions. We have given efficient algorithms for both the methods with provable guarantees for learning under label noise. The proposed algorithms are easy to implement and the classification performance is impressive even at high noise rates and competitive with state-of-the-art methods on benchmark data. The algorithms already give a new family of methods that can be applied to the positive-unlabeled learning problem [Elkan and Noto, 2008], but the implications of the methods for this setting should be carefully analysed. We could consider harder noise models such as label noise depending on the example, and "nasty label noise" where labels to flip are chosen adversarially.

7 Acknowledgments

This research was supported by DOD Army grant W911NF-10-1-0529 to ID; PR acknowledges the support of ARO via W911NF-12-1-0390 and NSF via IIS-1149803, IIS-1320894.

[3] A family of methods proposed by Crammer and coworkers [Crammer et al., 2006, 2009, Dredze et al., 2008] could be compared to, but [Crammer and Lee, 2010] show that the 2 NHERD variants perform the best.

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 *Proceedings of the 22nd Conference on Learning Theory*, 2009.

Battista Biggio, Blaine Nelson, and Pavel Laskov. Support vector machines under adversarial label noise. *Journal of Machine Learning Research - Proceedings Track*, 20:97–112, 2011.

Tom Bylander. Learning linear threshold functions in the presence of classification noise. In *Proc. of the 7th COLT*, pages 340–347, NY, USA, 1994. ACM.

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.

K. Crammer and D. Lee. Learning via gaussian herding. In *Advances in NIPS 23*, pages 451–459, 2010.

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

Koby Crammer, Alex Kulesza, and Mark Dredze. Adaptive regularization of weight vectors. In *Advances in NIPS 22*, pages 414–422, 2009.

Mark Dredze, Koby Crammer, and Fernando Pereira. Confidence-weighted linear classification. In *Proceedings of the Twenty-Fifth ICML*, pages 264–271, 2008.

C. Elkan and K. Noto. Learning classifiers from only positive and unlabeled data. In *Proc. of the 14th ACM SIGKDD intl. conf. on Knowledge discovery and data mining*, pages 213–220, 2008.

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

T. Graepel and R. Herbrich. The kernel Gibbs sampler. In *Advances in NIPS 13*, pages 514–520, 2000.

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

Neil D. Lawrence and Bernhard Schölkopf. Estimating a kernel Fisher discriminant in the presence of label noise. In *Proceedings of the Eighteenth 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. *To appear in IEEE Trans. Syst. Man and Cybern. Part B*, 2013. URL: `http://arxiv.org/abs/1109.5231`.

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.

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. *To appear in COLT*, 2013.

G. Stempfel and L. Ralaivola. Learning kernel perceptrons on noisy data using random projections. In *Algorithmic Learning Theory*, 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 *Proc. of the 19th Intl. Conf. on Artificial Neural Networks: Part I*, pages 884–893. Springer-Verlag, 2009.

Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the Twentieth ICML*, pages 928–936, 2003.

A Proofs

Proof of Lemma 1. 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

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

□

Proof of Lemma 2. By the basic Rademacher bound on the maximal deviation between risks and empirical risks over $f \in \mathcal{F}$, we get

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

where

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

If ℓ is L-Lipschitz then $\tilde{\ell}$ is L_ρ 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

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

□

Proof of Theorem 3. Let $f^\star$ be the minimizer of $R_{\ell, D}(\cdot)$ over $\mathcal{F}$. We have

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

We can now apply Lemma 2 to control the last quantity above, and thus obtain the first statement of the theorem. Now, if ℓ is *classification-calibrated*, then from Theorem 1 of [Bartlett et al., 2006], we know there exists a convex, invertible, nondecreasing transformation ψ_ℓ with $\psi_\ell(0) = 0$ such that,

$$\psi_\ell(R_D(f) - R^*) \leq R_{\ell, D}(f) - \inf_f R_{\ell, D}(f)$$

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

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

□

Proof of Lemma 4. 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 ℓ is convex in t. □

Proof of Lemma 5. 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. □

Proof of Theorem 6. The first part of the theorem follows by combining Lemma 2 and Lemma 5, using the fact that if $\|\mathbf{w}\|_2 \leq W_2$ for any $\mathbf{w}$ and $\|X_i\|_2 \leq X_2$ then, $\mathfrak{R}(\mathcal{W}) \leq W_2 X_2/\sqrt{n}$. The second part follows by noting that Theorem 3 is true also for 2ε-minimizers of the empirical risk $\widehat{R}_{\tilde{\ell}}$ provided we add 2ε to the right hand side. □

Proof of Lemma 7. The first equality is true because the optimal bayes classifier under D_ρ thresholds $\tilde{\eta}(X) = P(\tilde{Y} = 1|X)$ at 1/2. Now,

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

Therefore,

$$\begin{aligned}
\text{sign}(\tilde{\eta}(x) - 1/2) &= \text{sign}((1-\rho_{+1}-\rho_{-1})\eta(x) + \rho_{-1} - 1/2) \\
&= \text{sign}\left(\eta(x) - \frac{1/2 - \rho_{-1}}{1-\rho_{+1}-\rho_{-1}}\right).
\end{aligned}$$

□

Proof of Theorem 9. Let us think of f as $\{\pm 1\}$-valued since both C_D and C_{α,D_ρ} depend only on $\text{sign}(f)$. We have,

$$C_D(f) = \mathbb{E}_Y\left[1_{\{f(X)\neq Y\}}\right]$$

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 $R_D(f) = \mathbb{E}_X\left[C_D(f)\right]$, and $R_{\alpha,D_\rho}(f) = \mathbb{E}_X\left[C_{\alpha,D_\rho}(f)\right]$. Also note that $C_D(f) = \eta(X)$ if $f(X) = -1$, and $C_D(f) = 1 - \eta(X)$ 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:

$$\begin{aligned}
(1-\alpha)\tilde{\eta}(X) &= A\eta(X) + B \\
\alpha(1-\tilde{\eta}(X)) &= A(1-\eta(X)) + B
\end{aligned}$$

Using the relation between $\eta(X)$ and $\tilde{\eta}(X)$ in Lemma 7 and solving for A we get,

$$A = \frac{(1-\rho_{+1}-\rho_{-1})\eta(X) + \rho_{-1} - \alpha}{2\eta(X) - 1}.$$

Choosing $\alpha = \alpha^* = \frac{1-\rho_{+1}+\rho_{-1}}{2}$, and simplifying, we get a constant A that depends only on the noise rates:

$$A = A_\rho = \frac{1 - \rho_{+1} - \rho_{-1}}{2}.$$

Consequently,

$$B = \rho_{-1}(1 - \alpha^*) - \frac{\alpha^*}{2}(1 - \rho_{+1} - \rho_{-1})\eta(X).$$

Taking expectation with respect to X, we conclude:

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

where $B_X = \mathbb{E}_X[B]$. □

Proof of Corollary 10. The proof is immediate from Theorem 9 observing that B_X is independent of f. □

Proof of Theorem 11. From Corollary 4.1 in [Scott, 2012], we can infer that ℓ_α is α-CC for given $\alpha \in (0, 1)$, as ℓ is convex, classification-calibrated and $\ell^{'}(0) < 0$. Then, from Theorem 3.1 in [Scott, 2012], there exists an *invertible*, *non-decreasing* convex transformation ψ_{ℓ_α} 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_ρ, and let $f = \hat{f}_\alpha$. The RHS of the above inequality can then be controlled similarly as in the proof of Theorem 3. It is easy to see that the Lipschitz constant of ℓ_α is same as that of ℓ, 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_{α,D_ρ} and R_D at α^* (Theorem 9), we get $R_{\alpha^*,D_\rho}(f) - \min_f R_{\alpha^*,D_\rho}(f) = A_\rho(R_D(f) - R^*)$. B_X vanishes because it is constant for the distribution D_ρ. 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, the statement of the theorem follows: With probability at least $1 - \delta$,

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

□

B Online learning

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 $\mathbf{x}_i$ where $\tilde{y}_i$ are the noisy labels. 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 we choose $\lambda_1 \in \partial\ell(t, y)$ and $\lambda_2 \in \partial\ell(t, -y)$, (where $\partial\ell$ is the subdifferential w.r.t. t) we have

$$\mathbb{E}_{\tilde{y}}[g(t, \tilde{y})] \in \partial\ell(t, y) \tag{2}$$

where

$$g(t, y) = \frac{(1 - \rho_{-y})\lambda_1 - \rho_y\,\lambda_2}{1 - \rho_{+1} - \rho_{-1}} \tag{3}$$

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 (3)
end for

Theorem 12. *Let $\ell(t, y)$ be convex and L-Lipschitz in t (for every y). Fix an arbitrary sequence $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$. 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 ρ_{+1}, ρ_{-1}, then we have*

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

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 $\mathbf{w}$,

$$\sum_{i=1}^{n} \langle g_i, \mathbf{w}_{i-1} - \mathbf{w} \rangle \leq \frac{\gamma}{2} \sum_{i=1}^{n} \|g_i\|^2 + \frac{\|\mathbf{w}\|^2}{2\gamma}\,. \tag{4}$$

Since ℓ is L-Lipschitz, the λ_1, λ_2 appearing in the definition (3) 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

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

because $\mathbb{E}_{\tilde{y}_i}\left[g_i \mid \tilde{y}_{1:i-1}\right] \in \partial_{\mathbf{w}=\mathbf{w}_{i-1}} \ell(\langle \mathbf{w}, \mathbf{x}_i \rangle, y_i)$ by (2) and the chain rule for differentiation, and $\ell(\langle \mathbf{w}, \mathbf{x}_i \rangle, y_i)$ is convex in $\mathbf{w}$. Thus, for any $\mathbf{w}$ with $\|\mathbf{w}\|_2 \leq W_2$,

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

Since the above inequality is true for any $\mathbf{w}$ with $\|\mathbf{w}\|_2 \leq 1$, we have

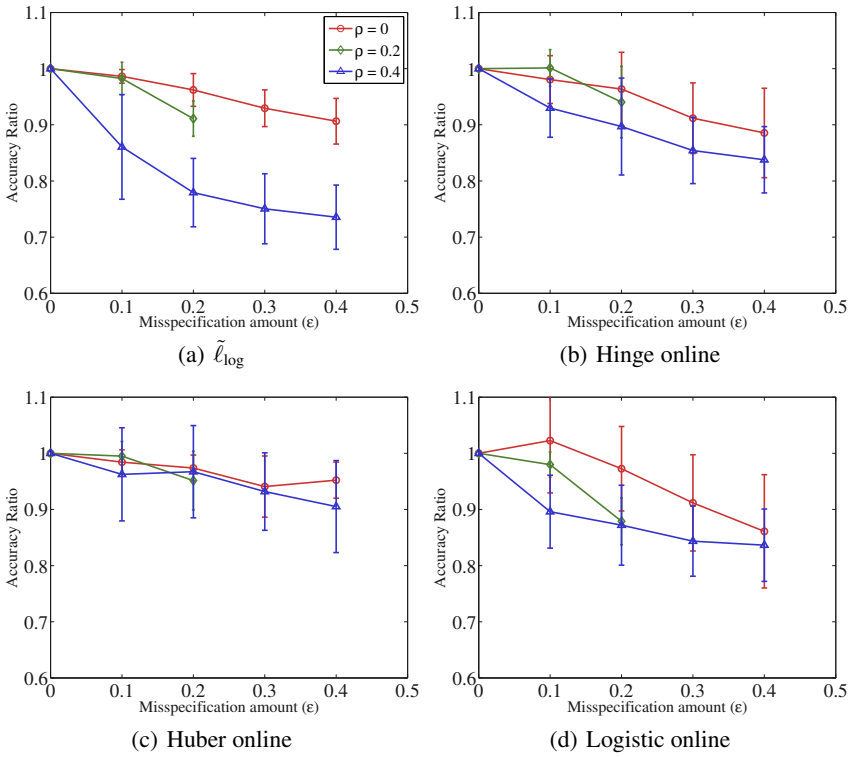
$$\mathbb{E}_{\tilde{y}_{1:n}}\left[\sum_{i=1}^{n} \ell(\langle \mathbf{w}_{i-1}, \mathbf{x}_i \rangle, y_i)\right] - \min_{\|\mathbf{w}\|_2 \leq W_2} \sum_{i=1}^{n} \ell(\langle \mathbf{w}, \mathbf{x}_i \rangle, y_i) \leq L_\rho X_2 W_2 \sqrt{n}.$$

Observing that the minimum over $\mathbf{w}$ is not random allows us to move it inside the expectation giving us the theorem. □

C Experiments

C.1 Knowledge of noise rates

The proposed algorithms require the knowledge of noise rates ρ_{+1} and ρ_{-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. We emphasize here that in case the true noise rates are known, our methods can benefit from that knowledge as observed from our experiments (results not shown), whereas the



(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 ρ_{+1} and ρ_{-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 ϵ and the accuracy at $\epsilon = 0$, i.e. when true noise rates are specified, is plotted for different values of noise rates ρ. 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 ϵ means that the performance of the algorithm, on average, is unaltered by misspecification of noise rates up to ϵ. As expected, the ratio decreases, i.e. the algorithms perform worse as ϵ increases. Most of the ratios being close to 1 suggests that the proposed methods are fairly robust with respect to ϵ-misspecification of noise rates.

competitive methods *cannot* as they do not involve noise rates. 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 ϵ and the accuracy at $\epsilon = 0$, i.e. when true noise rates are specified, is a measure of sensitivity of the algorithms to ϵ-misspecification of noise rates. We would want the ratio to be close to 1 for a given ϵ, which would suggest that the method is fairly robust with respect to the ϵ-misspecification. The results in Figure 3 show that the proposed methods are robust to ϵ-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.