Electronic Journal of Statistics
Vol. 0 (0000) 1
ISSN: 1935-7524
DOI: 10.1214/154957804100000000

Binary classification with corrupted labels

Yonghoon Lee and Rina Foygel Barber

Department of Statistics
The University of Chicago
e-mail: yhoony31@uchicago.edu; rina@uchicago.edu

Abstract: In a binary classification problem where the goal is to fit an accurate predictor, the presence of corrupted labels in the training data set may create an additional challenge. However, in settings where likelihood maximization is poorly behaved—for example, if positive and negative labels are perfectly separable—then a small fraction of corrupted labels can improve performance by ensuring robustness. In this work, we establish that in such settings, corruption acts as a form of regularization, and we compute precise upper bounds on estimation error in the presence of corruptions. Our results suggest that the presence of corrupted data points is beneficial only up to a small fraction of the total sample, scaling with the square root of the sample size.

MSC2020 subject classifications: Primary 62H30.
Keywords and phrases: Classification, Label noise.

1. Introduction

Consider a classification problem, where our goal is to predict a binary label $Y \in \{\pm 1\}$ using information captured by a feature vector $X \in \mathbb{R}^d$. Based on n training data points $(X_1, Y_1), \dots (X_n, Y_n)$, the objective is to fit a classifier $\widehat{f} : \mathbb{R}^d \to \{\pm 1\}$ to this data, mapping a new test feature vector X to a predicted label $+1$ or -1.

In many settings, inherent noise in the measurement process can introduce corruption into the observed labels Y_i. For example, consider a medical application where features X_i for patient i determine their likelihood of having a particular disease, and $Y_i \in \{\pm 1\}$ indicates presence or absence of the disease. Imperfect diagnostic tests might mean that the observed label may differ from the true label Y_i. Writing $\widetilde{Y}_i \in \{\pm 1\}$ to denote the observed label, we might have $\mathbb{P}\{\widetilde{Y}_i = -1 \mid Y_i = +1\} > 0$ (if the diagnostic test has a nonzero rate of false negatives) and similarly $\mathbb{P}\{\widetilde{Y}_i = +1 \mid Y_i = -1\} > 0$ (indicating false positives).

1.1. Setting and notation

We begin by introducing some basic notation and definitions that we will use throughout. Consider the following model for the triples $(X, Y, \widetilde{Y})$, where as before, $X \in \mathbb{R}^d$ denotes the feature vector, $Y \in \{\pm 1\}$ is the true label (which we do not observe), and $\widetilde{Y} \in \{\pm 1\}$ is the observed label (which may be corrupted, i.e., may differ from the true label):

$$\begin{aligned} X &\sim P_X \quad \text{(a distribution on } \mathbb{R}^d\text{)}, \\ Y|X &= \begin{cases} +1, & \text{with prob. } \eta(X), \\ -1, & \text{with prob. } 1 - \eta(X), \end{cases} \\ \widetilde{Y}|X, Y &= \begin{cases} -Y, & \text{with prob. } \rho, \\ Y, & \text{with prob. } 1 - \rho. \end{cases} \end{aligned}$$

Here $\eta(x)$ denotes the probability of a positive (true) label,

$$\eta(x) = \mathbb{P}\{Y = +1 \mid X = x\},$$

while ρ denotes the probability that the observed label is corrupted, assumed to be identical across all data points (the "homogeneous noise" setting).

In the classification problem, our goal is to define a classification rule that, given a feature vector $x \in \mathbb{R}^d$, outputs a predicted label $+1$ or -1. The misclassification rate is minimized by predicting $+1$ or -1 depending on whether $\eta(x)$ is above or below 0.5, respectively. In a real data setting where $\eta(x)$ is unknown, the classification problem is typically addressed by fitting some function $f(x) \in \mathbb{R}$ and then predicting the label $\text{sign}(f(x))$. We can interpret $f(x)$ as containing information about both our prediction for the label (via the sign) and our confidence in this prediction (via the magnitude—values $f(x) \approx 0$ indicate uncertainty).

Given a possible choice of the function f, the misclassification rate on the training data set $\{(X_i, Y_i) : i = 1, \ldots, n\}$ is therefore given by the empirical 0-1 loss,

$$\widehat{\mathcal{L}}_n^{0/1}(f) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\left\{f(X_i) \cdot Y_i \leqslant 0\right\},$$

while

$$\widetilde{\mathcal{L}}_n^{0/1}(f) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\left\{f(X_i) \cdot \widetilde{Y}_i \leqslant 0\right\}$$

measures misclassification on the *corrupted* training data set $\{(X_i, \widetilde{Y}_i) : i = 1, \ldots, n\}$. Our goal is to ensure a low "true" misclassification rate, i.e., for predicting the label Y for a new point with features X, that is,

$$\mathcal{L}^{0/1}(f) = \mathbb{P}\{f(X) \cdot Y \leqslant 0\},$$

where (X, Y) is a new data point drawn from the same distribution as the original training data—that is, $X \sim P_X$, and $Y|X$ is a label in $\{\pm 1\}$ with probabilities determined by $\eta(X)$.

Since the zero/one loss is challenging to optimize, it is standard to use a *surrogate loss function* $\ell : \mathbb{R} \to \mathbb{R}_+$, typically chosen to be continuous, convex, and monotone nonincreasing. For example, a logistic surrogate loss is given by

$$\ell(t) = \log(1 + e^{-t}),$$

while the hinge loss is given by

$$\ell(t) = \max\{0, 1 - t\}.$$

Given a sample of n data points, $(X_1, Y_1), \ldots, (X_n, Y_n)$, we then define the *empirical risk*

$$\widehat{\mathcal{L}}_n(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(f(X_i) \cdot Y_i),$$

which is the average surrogate loss on the data set $\{(X_i, Y_i) : i = 1, \ldots, n\}$, and the *corrupted empirical risk*

$$\widetilde{\mathcal{L}}_n^{\rho}(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(f(X_i) \cdot \widetilde{Y}_i),$$

which is the average surrogate loss on the *corrupted* data set $\{(X_i, \widetilde{Y}_i) : i = 1, \ldots, n\}$. We will also write

$$\mathcal{L}(f) = \mathbb{E}[\ell(f(X) \cdot Y)],$$

the "true" risk of a function f, with expectation taken over a data point (X, Y) drawn from the same distribution as before, i.e., $X \sim P_X$, and label $Y|X$ drawn with probabilities determined by $\eta(X)$.

1.2. Summary of questions and results

The key question of this work is to compare the performance of the empirical risk minimizer,

$$\widehat{f} = \text{argmin}_{f \in \mathcal{F}}\, \widehat{\mathcal{L}}_n(f),$$

and its corrupted counterpart,

$$\widetilde{f} = \operatorname{argmin}_{f \in \mathcal{F}} \widetilde{\mathcal{L}}_n^\rho(f),$$

where the minimization is taken over some predefined class of functions $\mathcal{F}$ (for example, linear functions of x). That is, how does the presence of corrupted labels affect the performance of the empirical risk minimizer? In particular, we emphasize that the surrogate loss function is unchanged—we do not adjust ℓ or attempt to "correct" for the presence of corruption (this is in contrast to much of the existing literature, which we review below).

Our findings can be summarized as follows. First, we find that **corruption mimics regularization**—in particular, for a fixed function $f \in \mathcal{F}$, the corrupted empirical risk $\widetilde{\mathcal{L}}_n^\rho(f)$ is a *biased* estimate of the true risk $\mathcal{L}(f)$, but acts as an *unbiased* estimate of a penalized version of this risk,

$$\mathcal{L}(f) + \lambda \mathsf{R}(f)$$

where $\lambda > 0$ is a penalty parameter depending on the corruption level ρ, while the regularization function is given by

$$\mathsf{R}(f) = \mathbb{E}\left[\frac{\ell(f(X)) + \ell(-f(X))}{2}\right],$$

the expected loss of the function f under a completely random label.

While adding a penalty introduces bias into our estimator, it also serves to reduce variance, and for limited sample size n, this reduction in variance may outweigh the bias. Our second finding is therefore that, in some settings, **corruption may lead to reduced risk for finite sample size**, since it is effectively acting as a regularizer and can substantially reduce variance.

1.3. Prior work

The problem of learning a classifier in the presence of corrupted labels has been studied in many works in the recent literature. Here we give a very brief overview of the settings and types of results considered. Consider the more general model

$$\begin{aligned} X &\sim P_X \quad \text{(a distribution on } \mathbb{R}^d\text{)}, \\ Y|X &= \begin{cases} +1, & \text{with prob. } \eta(X), \\ -1, & \text{with prob. } 1-\eta(X), \end{cases} \\ \widetilde{Y}|X,Y &= \begin{cases} -Y, & \text{with prob. } \rho(X,Y), \\ Y, & \text{with prob. } 1-\rho(X,Y). \end{cases} \end{aligned}$$

Here $\eta(x)$ denotes the probability of a positive (true) label as before, while $\rho(x,y)$ denotes the probability that the observed label is corrupted,

$$\rho(x,y) = \mathbb{P}\{\widetilde{Y} \neq Y \mid X = x, Y = y\},$$

which now may depend on x and/or y.

Frénay et al. [7] and Frenay and Verleysen [8] provide overviews of recent works on this problem. They categorize the existing methods to three types: label noise-robust models, data cleaning methods, and label noise-tolerant learning algorithms.

The *homogeneous noise* setting assumes that $\rho(x,y) \equiv \rho$ for all x, y—that is, there is a constant probability for each label to be corrupted. This is the setting we study in the present work. Under this setting, Long and Servedio [14] study boosting algorithms and discuss negative consequences of label noise. Van Rooyen, Menon and Williamson [26] consider ERM method and propose a label noise-robust loss function. Manwani and Sastry [16] discuss the noise-tolerance property of risk minimization. Cannings, Fan and Samworth [5] show that LDA is consistent under the noise, and Blanco, Japón and Puerto [2] propose robust algorithms that apply relabeling and clustering to SVM.

The *class-dependent noise* setting assumes that $\rho(x, y) = \rho_y$ for all x, y—that is, the probability of corrupting a positive label ($Y = +1$ but $\widetilde{Y} = -1$) is constant with respect to the feature vector x, and similarly for a negative label, but these two probabilities may differ. For example, in our earlier medical example, the diagnostic test might have different false positive and false negative rates, but these rates themselves are constant across patients (i.e., independent of features such as age that might be included in the X vector). Liu and Tao [13], Scott, Blanchard and Handy [25], and Blanchard et al. [1] study the consistency of the classifier under corruption, while Reeve et al. [23] focus on the minimax optimal learning rate of the corrupted estimator. Some recent works try correction of the loss function or the observed labels; see Natarajan et al. [19], van Rooyen and Williamson [27], Patrini et al. [21], and Lin and Bradic [12]. Other recent works focus on studying or developing label noise-robust methods; see Natarajan et al. [18], Patrini et al. [20], Reeve and Kabán [24], Bootkrajang and Kabán [3], and Bootkrajang and Kabán [4].

Finally, the *general* setting—where $\rho(x, y)$ might vary with x—is studied by Cannings, Fan and Samworth [5]. In particular, they examine a setting for k-nearest neighbor where the corrupted labels $\widetilde{Y}_i$ are more "clean" than the original labels Y_i, in the sense that the corruption mechanism defined by $\rho(x, y)$ acts to denoise labels near the decision boundary (i.e., $\eta(x) \approx 0.5$) Specifically, suppose that, for values x with $\eta(x)$ slightly higher than 0.5, we have $\rho(x, +1) < \rho(x, -1)$ (that is, a label $Y_i = -1$ that "should" instead be positive, has a greater chance of being flipped to $\widetilde{Y}_i = +1$), and similarly if $\eta(x)$ is slightly lower than 0.5 then $\rho(x, +1) > \rho(x, -1)$. In this case, the $\widetilde{Y}_i$'s carry strictly more information for estimating the decision boundary, as compared to the Y_i's; this setting is therefore fundamentally different from the one we consider here, where homogeneous noise creates strictly noisier labels. Menon, Van Rooyen and Natarajan [17] consider a similar general setting where they show that any consistent algorithm for noise free setting is also consistent under noisy labels under appropriate assumptions. Recent discussions on the noise-tolerence and the robustness of the corrupted classification under this setting can be found in Ghosh, Manwani and Sastry [9] and Cheng et al. [6].

2. Main results

2.1. Intuition: corruption acts as regularization

The key idea for studying the corrupted estimator through the framework of regularization, is to find a regularizer that matches the expected behavior of the corruption. In order to do this, we first find a different representation of the corruption variables: define

$$R_i \overset{\text{iid}}{\sim} \text{Bernoulli}(2\rho) \text{ and } Z_i \overset{\text{iid}}{\sim} \text{Uniform}\{\pm 1\},$$

drawn independently from each other and independently of the clean data. Then let

$$\widetilde{Y}_i = (1 - R_i) \cdot Y_i + R_i \cdot Z_i.$$

That is, R_i determines whether the label Y_i will be replaced by a random sign, and Z_i provides this random sign. Examining this construction we can see that this yields the same distribution of the corrupted labels as the original definition. We can then write the corrupted loss as

$$\widetilde{\mathcal{L}}_n^{\rho}(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(f(X_i) \cdot \widetilde{Y}_i) = \frac{1}{n}\sum_{i=1}^{n}(1 - R_i) \cdot \ell(f(X_i) \cdot Y_i) + \sum_{i=1}^{n} R_i \cdot \ell(f(X_i) \cdot Z_i).$$

Next, we treat f as fixed, and then condition on the clean data and marginalize over the distribution of the R_i's and Z_i's:

$$\begin{aligned}\mathbb{E}\left[\widetilde{\mathcal{L}}_n^\rho(f) \;\middle|\; X_{1:n}, Y_{1:n}\right] &= \frac{1}{n}\sum_{i=1}^n \mathbb{E}[1-R_i]\cdot\ell(f(X_i)\cdot Y_i) + \frac{1}{n}\sum_{i=1}^n \mathbb{E}[R_i]\cdot\mathbb{E}[\ell(f(X_i)\cdot Z_i) \mid X_i] \\ &= (1-2\rho)\cdot\widehat{\mathcal{L}}_n(f) + \rho\cdot\frac{1}{n}\sum_{i=1}^n\big(\ell(f(X_i)) + \ell(-f(X_i))\big).\end{aligned}$$

Recall the definition of the regularizer,

$$\mathsf{R}(f) = \mathbb{E}\left[\frac{\ell(f(X)) + \ell(-f(X))}{2}\right],$$

the expected loss of f on purely random labels. We can also consider an empirical version,

$$\widehat{\mathsf{R}}_n(f) = \frac{1}{n}\sum_{i=1}^n \frac{\ell(f(X_i)) + \ell(-f(X_i))}{2}.$$

We therefore see that

$$\mathbb{E}\left[\widetilde{\mathcal{L}}_n^\rho(f) \;\middle|\; (X_i, Y_i), i = 1,\ldots,n\right] = (1-2\rho)\cdot\Big(\widehat{\mathcal{L}}_n(f) + \lambda\widehat{\mathsf{R}}_n(f)\Big),$$

where $\lambda = \frac{2\rho}{1-2\rho}$. Finally, for any fixed function f, we have

$$\mathbb{E}[\widehat{\mathcal{L}}_n(f) + \lambda\widehat{\mathsf{R}}_n(f)] = \mathcal{L}(f) + \lambda\mathsf{R}(f),$$

by definition. Therefore, we can view the corrupted empirical risk minimizer $\widetilde{f}$ as a sample estimate of the minimizer of the penalized loss $\mathcal{L}(f) + \lambda\mathsf{R}(f)$.

To summarize our findings so far, we have seen that $\widetilde{f} = \operatorname{argmin}_{f\in\mathcal{F}}\widetilde{\mathcal{L}}_n^\rho(f)$ can be described in two ways:

- Fixing the training data $\{(X_i, Y_i) : i = 1,\ldots,n\}$ and taking an expectation over the corruption mechanism (the R_i's and Z_i's above), we see that $\widetilde{\mathcal{L}}_n^\rho(f)$ has (conditional) expected value $\widehat{\mathcal{L}}_n(f) + \lambda\widehat{\mathsf{R}}_n(f)$, a penalized empirical risk.
- Taking expectations over both the original data and the random corruption, $\widetilde{\mathcal{L}}_n^\rho(f)$ has expected value $\mathcal{L}(f) + \lambda\mathsf{R}(f)$, a penalized true risk.

2.2. Results for the linear setting

Next, we will examine the implications of this relationship between corruption and regularization, on the goals of minimizing risk. From this point on, we will restrict our discussion to the setting where $\mathcal{F}$ consists of *linear* functions,

$$\mathcal{F} = \{x \mapsto w^\top x : w \in \mathbb{R}^d\},$$

in order to be able to achieve precise results. Consequently we will shift our notation from functions f to vectors w. Specifically, for each $w \in \mathbb{R}^d$ we will define the population-level loss and regularized loss,

$$\mathcal{L}(w) = \mathbb{E}[\ell(X^\top w\cdot Y)] \quad\text{and}\quad \widetilde{\mathcal{L}}^\rho(w) = \mathbb{E}[\ell(X^\top w\cdot Y)] + \frac{2\rho}{1-2\rho}\cdot\mathsf{R}(w),$$

where

$$\mathsf{R}(w) = \mathbb{E}\left[\frac{\ell(X^\top w) + \ell(-X^\top w)}{2}\right] = \frac{\mathcal{L}(w) + \mathcal{L}(-w)}{2},$$

as well as the empirical loss and empirical corrupted loss,

$$\widehat{\mathcal{L}}_n(w) = \frac{1}{n}\sum_{i=1}^{n} \ell(X_i^\top w \cdot Y_i) \quad \text{and} \quad \widetilde{\mathcal{L}}_n^\rho(w) = \frac{1}{n}\sum_{i=1}^{n} \ell(X_i^\top w \cdot \widetilde{Y}_i).$$

We will also define population-level minimizers

$$w_* = \text{argmin}_{w\in\mathbb{R}^d} \mathcal{L}(w) \quad \text{and} \quad \widetilde{w}_*^\rho = \text{argmin}_{w\in\mathbb{R}^d} \widetilde{\mathcal{L}}^\rho(w), \tag{1}$$

and empirical minimizers

$$\widehat{w}_n = \text{argmin}_{w\in\mathbb{R}^d} \widehat{\mathcal{L}}_n(w) \quad \text{and} \quad \widetilde{w}_n^\rho = \text{argmin}_{w\in\mathbb{R}^d} \widetilde{\mathcal{L}}_n^\rho(w), \tag{2}$$

whenever these minimizers exist. (Note that, in some settings, the loss or its empirical or corrupted counterpart may have no minimizer—for example, logistic loss, where the positive and negative labels can be perfectly separated.) For each of the four minimization problems, if the minimizer exists but is not unique, our results will apply to any minimizer (e.g., $\widetilde{w}_*^\rho$ denotes any element of the set $\text{argmin}_{w\in\mathbb{R}^d} \widetilde{\mathcal{L}}^\rho(w)$, etc).

It is well-known that regularization may help reduce risk, even at the cost of increasing bias due to the influence of the regularization function. As discussed earlier, since corruption mimics regularization, in many settings we empirically observe that corruption reduces the risk—that is, $\mathcal{L}(\widetilde{w}_n^\rho) < \mathcal{L}(\widehat{w}_n)$, even though the corruption introduces bias. We will next study why this phenomenon occurs, by establishing bounds on the loss $\mathcal{L}(\widetilde{w}_n^\rho)$ of the corrupted estimator.

2.2.1. Theoretical results

We begin by defining our assumptions. First, we require some conditions on the loss function ℓ:

Assumption 1. *The loss function ℓ is nonnegative, nonincreasing, convex, and L-Lipschitz. Furthermore, ℓ is strictly decreasing on negative values, with*

$$\ell(t) \geqslant \ell(0) + \gamma|t| \text{ for all } t \leqslant 0$$

for some $\gamma > 0$, and has a subexponential decay for positive values,

$$\ell(t) \leqslant c_1 e^{-c_2 t} \text{ for all } t \geqslant 0,$$

for some $c_1, c_2 > 0$.

The last two conditions ensure that the loss function enacts a strong penalty if $X^\top w$ predicts the sign of Y incorrectly (i.e., $\ell(t)$ is large for $t < 0$), but decays quickly if $X^\top w$ predicts the sign of Y correctly (i.e., $\ell(t)$ is small for $t > 0$). These conditions are satisfied by many well-known examples, for instance:

- The logistic loss $\ell_t = \log(1 + e^{-t})$ satisfies Assumption 1 with $\gamma = \frac{1}{2}$ and $L = c_1 = c_2 = 1$.
- The hinge loss $\ell_t = (1 - t)_+$ satisfies Assumption 1 with $L = \gamma = c_1 = c_2 = 1$.

We will also need some weak assumptions on the distribution of the feature vector X:

Assumption 2. *For some $a_0, a_1, a_2 > 0$, it holds that*

$$\mathbb{E}[e^{a_0|X^\top u|^2}] \leqslant a_1$$

and

$$\mathbb{E}\left[e^{-t|X^\top u|}\right] \leqslant \frac{a_2}{t} \text{ for all } t > 0.$$

for all unit vectors $u \in \mathbb{S}^{d-1}$.

For example, this assumption is satisfied by any multivariate Gaussian distribution with mean μ and covariance Σ, with the parameters a_0, a_1, a_2 depending on $\|\mu\|$ and on the largest and smallest eigenvalues of Σ, but not on the dimension d.

Under these assumptions, our main result establishes a bound on the loss of the corrupted estimator $\widetilde{w}_n^\rho$.

Theorem 1. *Suppose that Assumptions 1 and 2 hold. Let $n \geqslant 2$ and fix any $\alpha > 0$. Suppose $\rho \in (0, \frac{1}{2})$ satisfies*

$$\rho \geqslant C \cdot \frac{d \log n}{n}.$$

Then with probability at least $1-n^{-\alpha}$, the set $\operatorname{argmin}_{w\in\mathbb{R}^d} \widetilde{\mathcal{L}}_n^\rho(w)$ is nonempty, and for all $\widetilde{w}_n^\rho \in \operatorname{argmin}_{w\in\mathbb{R}^d} \widetilde{\mathcal{L}}_n^\rho(w)$ it holds that

$$\mathcal{L}(\widetilde{w}_n^\rho) \leqslant \inf_{w\in\mathbb{R}^d} \mathcal{L}(w) + C' \left[\rho^{1/2} + \rho^{-1/2} \cdot \sqrt{\frac{d \log n}{n}} \right].$$

Here C, C' depend only on α and on the constants in Assumptions 1 and 2, but not on n, d, or ρ.

We can see an immediate tradeoff in the upper bound in Theorem 1. The $\rho^{1/2}$ term acts as an "approximation error", where a large corruption proportion ρ leads to a potentially large gap between the loss of the regularized estimator, $\mathcal{L}(\widetilde{w}_*^\rho)$, and the minimum possible loss without regularization, $\inf_{w\in\mathbb{R}^d} \mathcal{L}(w)$. On the other hand, the $\rho^{-1/2} \cdot \sqrt{\frac{d \log n}{n}}$ term is the "estimation error", which is large when the corruption proportion ρ is small (i.e., insufficient regularization). The resulting upper bound on risk is minimized when the corruption level scales as $\rho \asymp \left(\frac{d \log n}{n}\right)^{1/2}$, leading to an upper bound on excess risk scaling as $\asymp \left(\frac{d \log n}{n}\right)^{1/4}$. This suggests that even a very small fraction of corrupted entries can lead to a reduced risk. In contrast, the uncorrupted minimization problem may not behave well under these weak assumptions—for instance, if the labels are perfectly linearly separable (as might be the case if, e.g., $Y|X$ follows a logistic regression with very high signal strength), then a minimizer does not even exist (i.e., $\operatorname{argmin}_{w\in\mathbb{R}^d} \widehat{\mathcal{L}}_n(w)$ is empty).

The assumption that $\rho \geqslant C \cdot \frac{d \log n}{n}$ is not merely an artifact of the proof—in fact, without this type of assumption, we cannot even ensure that $\operatorname{argmin}_{w\in\mathbb{R}^d} \widetilde{\mathcal{L}}_n^\rho(w)$ is nonempty. To see why, let us consider a setting where the population is perfectly separable and ℓ is a strictly decreasing function. In this case, the empirical risk minimizer $\widehat{w}_n$ does not exist (or in other words, it diverges). Now, if $\rho = 1/n$, then with probability $(1 - \frac{1}{n})^n \approx e^{-1}$, the corrupted dataset is equal to the original dataset, which means that the corrupted data set is also perfectly separable and thus $\widetilde{w}_n^\rho$ does not exist.

Of course, the result of Theorem 1 is an upper bound on the loss, and may be loose for certain examples; the value of ρ that minimizes the upper bound (i.e., $\rho \asymp \left(\frac{d \log n}{n}\right)^{1/2}$) might not be the same as the value of ρ that minimizes the loss itself. In particular, the result can be viewed as a "worst case" bound that holds even when the unregularized loss has no minimizer (such as logistic regression with perfectly separable labels, as mentioned above); in problems where this is not the case, regularization is not as critical, and a smaller value of ρ (or even $\rho = 0$) may perform better.

2.2.2. Proof of Theorem 1

Our first step is to examine some properties of the regularized population minimizer $\widetilde{w}_*^\rho$ and its empirical counterpart, the corrupted estimator $\widetilde{w}_n^\rho$.

Lemma 1. *Suppose Assumptions 1 and 2 hold. Fix any $\rho \in (0, \frac{1}{2})$. Then $\operatorname{argmin}_{w\in\mathbb{R}^d} \widetilde{\mathcal{L}}^\rho(w)$ is nonempty, and any $\widetilde{w}_*^\rho \in \operatorname{argmin}_{w\in\mathbb{R}^d} \widetilde{\mathcal{L}}^\rho(w)$ must satisfy $\|\widetilde{w}_*^\rho\| \leqslant C_0 \rho^{-1/2}$ and*

$$\mathcal{L}(\widetilde{w}_*^\rho) \leqslant \inf_{w\in\mathbb{R}^d} \mathcal{L}(w) + C_1 \rho^{1/2}.$$

Moreover, for any $\alpha > 0$, if $n \geqslant 2$ and $\rho \geqslant C \cdot \frac{d \log n}{n}$ then with probability at least $1 - n^{-\alpha}$ it holds that $\operatorname{argmin}_{w \in \mathbb{R}^d} \widetilde{\mathcal{L}}_n^\rho(w)$ *is nonempty, that any* $\widetilde{w}_n^\rho \in \operatorname{argmin}_{w \in \mathbb{R}^d} \widetilde{\mathcal{L}}_n^\rho(w)$ *must satisfy* $\|\widetilde{w}_n^\rho\| \leqslant C_0 \rho^{-1/2}$*, and that*

$$\sup_{\|w\| \leqslant C_0 \rho^{-1/2}} \left| \widetilde{\mathcal{L}}_n^\rho(w) - \widetilde{\mathcal{L}}^\rho(w) \right| \leqslant C_2 \rho^{-1/2} \sqrt{\frac{d \log n}{n}}.$$

Here C, C_0, C_1, C_2 depend on α and on the constants in Assumptions 1 and 2, but not on n, d, or ρ.

Now we prove the theorem. By Lemma 1, with probability at least $1 - n^{-\alpha}$, for any $\widetilde{w}_*^\rho \in \operatorname{argmin}_{w \in \mathbb{R}^d} \widetilde{\mathcal{L}}^\rho(w)$ and all $\widetilde{w}_n^\rho \in \operatorname{argmin}_{w \in \mathbb{R}^d} \widetilde{\mathcal{L}}_n^\rho(w)$ it holds that $\mathcal{L}(\widetilde{w}_*^\rho) \leqslant \inf_{w \in \mathbb{R}^d} \mathcal{L}(w) + C_1 \rho^{1/2}$ and that

$$\max\left\{ \left| \widetilde{\mathcal{L}}_n^\rho(\widetilde{w}_*^\rho) - \widetilde{\mathcal{L}}^\rho(\widetilde{w}_*^\rho) \right|, \left| \widetilde{\mathcal{L}}_n^\rho(\widetilde{w}_n^\rho) - \widetilde{\mathcal{L}}^\rho(\widetilde{w}_n^\rho) \right| \right\} \leqslant C_2 \rho^{-1/2} \sqrt{\frac{d \log n}{n}}.$$

From now on, we assume that these events all hold. Then we have

$$\begin{aligned}
\widetilde{\mathcal{L}}^\rho(\widetilde{w}_n^\rho) &= \widetilde{\mathcal{L}}^\rho(\widetilde{w}_*^\rho) + \left(\widetilde{\mathcal{L}}_n^\rho(\widetilde{w}_*^\rho) - \widetilde{\mathcal{L}}^\rho(\widetilde{w}_*^\rho) \right) + \left(\widetilde{\mathcal{L}}_n^\rho(\widetilde{w}_n^\rho) - \widetilde{\mathcal{L}}_n^\rho(\widetilde{w}_*^\rho) \right) + \left(\widetilde{\mathcal{L}}^\rho(\widetilde{w}_n^\rho) - \widetilde{\mathcal{L}}_n^\rho(\widetilde{w}_n^\rho) \right) \\
&\leqslant \widetilde{\mathcal{L}}^\rho(\widetilde{w}_*^\rho) + \left(\widetilde{\mathcal{L}}_n^\rho(\widetilde{w}_n^\rho) - \widetilde{\mathcal{L}}_n^\rho(\widetilde{w}_*^\rho) \right) + 2C_2 \rho^{-1/2} \cdot \sqrt{\frac{d \log n}{n}} \\
&\leqslant \widetilde{\mathcal{L}}^\rho(\widetilde{w}_*^\rho) + 2C_2 \rho^{-1/2} \cdot \sqrt{\frac{d \log n}{n}} \quad \text{by optimality of } \widetilde{w}_n^\rho \\
&\leqslant \inf_{w \in \mathbb{R}^d} \mathcal{L}(w) + C_1 \rho^{1/2} + 2C_2 \rho^{-1/2} \cdot \sqrt{\frac{d \log n}{n}} \\
&\leqslant \inf_{w \in \mathbb{R}^d} \mathcal{L}(w) + \frac{C'}{2} \left[\rho^{1/2} + \rho^{-1/2} \cdot \sqrt{\frac{d \log n}{n}} \right],
\end{aligned}$$

where we set $C' = \max\{2C_1, 4C_2\}$. Next, by definition of $\widetilde{\mathcal{L}}^\rho$, we have

$$\begin{aligned}
\widetilde{\mathcal{L}}^\rho(\widetilde{w}_n^\rho) - \inf_{w \in \mathbb{R}^d} \mathcal{L}(w) &= (1 - \rho) \cdot \left[\mathcal{L}(\widetilde{w}_n^\rho) - \inf_{w \in \mathbb{R}^d} \mathcal{L}(w) \right] + \rho \cdot \left[\mathcal{L}(-\widetilde{w}_n^\rho) - \inf_{w \in \mathbb{R}^d} \mathcal{L}(w) \right] \\
&\geqslant \frac{1}{2} \left[\mathcal{L}(\widetilde{w}_n^\rho) - \inf_{w \in \mathbb{R}^d} \mathcal{L}(w) \right]
\end{aligned}$$

where the last step holds since $\rho \leqslant \frac{1}{2}$. Therefore,

$$\mathcal{L}(\widetilde{w}_n^\rho) \leqslant \inf_{w \in \mathbb{R}^d} \mathcal{L}(w) + C' \left[\rho^{1/2} + \rho^{-1/2} \cdot \sqrt{\frac{d \log n}{n}} \right],$$

which completes the proof of the theorem.

2.2.3. Another perspective on the regularizer

The results above suggest that the main source of possible improvements by corruption is the shrinkage induced by the corruption (or, at the population level, by the regularizer $\mathsf{R}(w)$). In particular, the results of Lemma 1 show that, in the linear setting, the corruption (or the regularizer) lead to an upper bound on $\|w\|$. We will now examine this connection more closely.

The following lemma verifies that, up to constants, $\mathsf{R}(w)$ is equivalent to $\|w\|$. In a sense, then, we can view regularization with $\mathsf{R}(w)$ as effectively placing a penalty on $\|w\|$.

Lemma 2. *Suppose Assumptions 1 and 2 hold. Then it holds that*

$$\max\{c_L \cdot \|w\|, \ell(0)\} \leqslant \mathsf{R}(w) \leqslant c_U \cdot \|w\| + \ell(0) \text{ for all } w \in \mathbb{R}^d,$$

where c_L, c_U depend only on the constants in Assumptions 1 and 2.

Proof. In the calculations (3) and (4) appearing in the proof of Lemma 1, we will see that Assumption 2 implies that

$$\frac{\log 2}{2a_2} \leqslant \mathbb{E}[|X^\top u|] \leqslant \sqrt{\frac{a_1}{a_0}}$$

for all unit vectors $u \in \mathbb{R}^d$. For any $w \in \mathbb{R}^d$, for the lower bound, we have

$$\begin{aligned}\mathsf{R}(w) = \mathbb{E}\left[\frac{\ell(|X^\top w|) + \ell(-|X^\top w|)}{2}\right] \geqslant \mathbb{E}\left[\frac{\ell(-|X^\top w|)}{2}\right] \geqslant \mathbb{E}\left[\frac{\ell(-|X^\top w|) - \ell(0)}{2}\right] \\ \geqslant \frac{\gamma}{2} \cdot \mathbb{E}[|X^\top w|] \geqslant \frac{\gamma \log 2}{4a_2} \cdot \|w\|,\end{aligned}$$

and furthermore

$$\mathsf{R}(w) = \mathbb{E}\left[\frac{\ell(|X^\top w|) + \ell(-|X^\top w|)}{2}\right] \geqslant \ell(0)$$

by convexity of ℓ. For the upper bound, we have

$$\begin{aligned}\mathsf{R}(w) &= \mathbb{E}\left[\frac{\ell(|X^\top w|) + \ell(-|X^\top w|)}{2}\right] \\ &= \ell(0) + \mathbb{E}\left[\frac{\ell(-|X^\top w|) - \ell(0)}{2}\right] + \mathbb{E}\left[\frac{\ell(|X^\top w|) - \ell(0)}{2}\right] \\ &\leqslant \ell(0) + \mathbb{E}\left[\frac{\ell(-|X^\top w|) - \ell(0)}{2}\right] \leqslant \ell(0) + \frac{L}{2} \cdot \mathbb{E}[|X^\top w|] \leqslant \ell(0) + \frac{L}{2}\sqrt{\frac{a_1}{a_0}} \cdot \|w\|.\end{aligned}$$

□

3. Simulations

Now we empirically investigate the effect of corruption through a simulation.[1] We generate the data $\{(X_i, Y_i)\}_{1 \leqslant i \leqslant n}$ in the following way: choosing dimension $d = 50$, we draw

$$\begin{aligned}X_i &\sim \mathcal{N}(0, \mathbf{I}_d) \\ Y_i \mid X_i &= \begin{cases} +1, & \text{with probability } \frac{\exp\{3X_{i1} + 0.5(X_{i2})^3\}}{1+\exp\{3X_{i1} + 0.5(X_{i2})^3\}}, \\ -1, & \text{with probability } \frac{1}{1+\exp\{3X_{i1} + 0.5(X_{i2})^3\}}, \end{cases}\end{aligned}$$

independently for each $i = 1, \ldots, n$. The corrupted labels $\{\widetilde{Y}_i\}_{1 \leqslant i \leqslant n}$ are generated as

$$\widetilde{Y}_i \mid X_i, Y_i = \begin{cases} -Y_i, & \text{with prob. } \rho, \\ Y_i, & \text{with prob. } 1 - \rho, \end{cases}$$

independently for each $i = 1, \ldots, n$. We run the experiment at a small and large sample size, $n = 400$ and $n = 2000$, and at a range of values of the corruption probability, $\rho \in \{0, 0.01, 0.02, \ldots, 0.2\}$. For each sample size n and corruption level ρ, we run 100 independent trials of the experiment, we choose the logistic loss function $\ell(t) = \log(1+e^{-t})$, and compute the corrupted empirical minimizer $\widetilde{w}_n^\rho$ defined in (2) and the penalized population-level minimizer $\widetilde{w}_*^\rho$ as in (1) (which reduces to the uncorrupted empirical minimizer $\widehat{w}_n$ and the unpenalized population-level minimizer w_*, respectively, in the case $\rho = 0$). Note that the data generating distribution does not follow the logistic regression model (due to the cubic term), and so the logistic loss simply acts as a surrogate for the 0-1 loss (i.e., it does not correspond to a likelihood for some well-specified model).

[1] Code to reproduce this simulation is available at https://www.stat.uchicago.edu/~rina/code/corrupted_labels_sim.R.

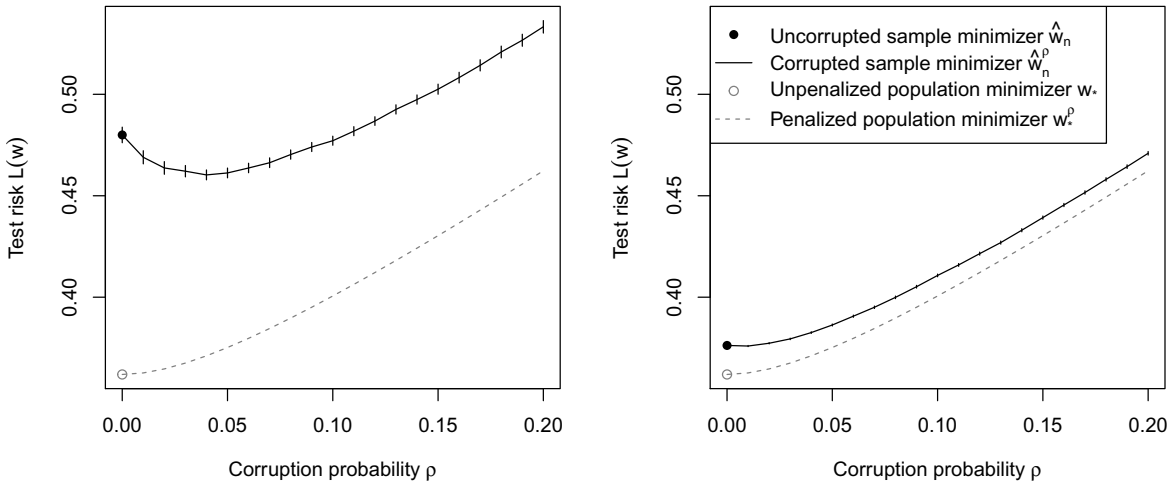

Fig 1: Risks of the original classifier $\widehat{w}_n$, the corrupted classifier $\widetilde{w}_n^\rho$, the optimal classifier w_*, and the population-level corrupted classifier $\widetilde{w}_*^\rho$ on the test set, with sample size $n = 400$ (left) and $n = 2000$ (right). For the sample estimators $\widehat{w}_n$ and $\widetilde{w}_n^\rho$, the figure displays the mean over 100 independent trials, with standard error bars. See Section 3 for further details.

Figure 1 shows the performance of the corrupted estimator $\widetilde{w}_n^\rho$ and its population-level version $\widetilde{w}_*^\rho$, across the range of corruption values $\rho \in \{0, 0.01, 0.02, \ldots, 0.2\}$, at each sample size $n \in \{400, 2000\}$; the result at $\rho = 0$ is highlighted in each case, as it corresponds to the uncorrupted estimator $\widehat{w}_n$ and to the corresponding population-level minimizer w_*. Overall, the plots illustrate how corruption acts as regularization—for the smaller sample size $n = 400$, we see that a small amount of corruption substantially reduces the test risk of the empirical minimizer $\widetilde{w}_n^\rho$, while for the larger sample size $n = 2000$ the uncorrupted estimator $\widehat{w}_n$ achieves good performance and we no longer see any noticeable improvement from corruption. For the population-level minimizers, on the other hand, increasing regularization always leads to an increase in risk, as expected.

4. Discussion

In this work, we have shown that the corruption of labels has a regularization-type effect on binary classification problems, leading to a possibility of an improvement of the fitted classifier in terms of test risk. Unlike many prior works that apply adjustment or correction to achieve consistency or robustness of the estimator, our result implies that corruption itself can be beneficial without any adjustment to the estimation process, and thus it could be better in some cases to simply fit the corrupted dataset without any modification on the methods—in particular, this means that we do not need to know or estimate the corruption mechanism, as would be the case for a procedure that corrects for the corruption. For the fitting of linear classifiers using empirical risk minimization under homogeneous noise, Theorem 1 provides an explanation for the possibility of corruption being beneficial, illustrating the tradeoff between loss approximation and the estimation.

We can expect a similar tradeoff for more general settings where the noise is not homogeneous, or where different estimation methods are applied; in general, it is intuitive that a small amount of corruption can reduce the chance of overfitting, especially when the inherent noise level is low, and that this benefit may outweigh the low bias that is introduced. As an example of a broader setting where this type of phenomenon may be useful, we can consider a setting where some data points are known to be "clean"

while others are potentially corrupted (this setting can be thought of as a special case of transfer learning—for example, see Reeve, Cannings and Samworth [22]). While we might expect that performance could be improved by removing or down-weighting the latter data points in order to avoid or reduce the effect of corruption, our findings instead suggest that the presence of the non-"clean" data might even be beneficial.

The question of corrupted labels, with its possible risks and benefits, is studied only in a very specific setting in our work (i.e., linear prediction rules in low dimensions), and many open questions remain. First, noting that the corrupted loss can be thought as another surrogate of 0-1 loss, we may ask how corruption affects the prediction performance of the estimator in terms of misclassification rate, i.e., 0-1 risk. Second, do similar phenomena occur in the high-dimensional regime, $d \gg n$ or $d \propto n$? In particular, we have seen that homogeneous corruption mimics an ℓ_2 penalty in the low-dimensional setting; however, the same is not immediately true in high dimensions, since these results rely on concentration type arguments that would no longer hold (and, in particular, for $d \gg n$, in general both the uncorrupted data $\{(X_i, Y_i)\}_{1 \leqslant i \leqslant n}$ and the corrupted data $\{(X_i, \widetilde{Y}_i)\}_{1 \leqslant i \leqslant n}$ are perfectly linearly separable, so we cannot expect good performance without some additional constraints or regularization). Finally, since the key phenomenon underlying our results is the way that homogeneous corruption mimics ℓ_2 regularization (and therefore, corruption induces shrinkage in the resulting estimator), this does not explain any potential benefits from corruption if we instead use methods such as a k-nearest-neighbor estimator, or other methods where there is no notion of shrinkage; is corruption beneficial more broadly, by reducing the chance of overfitting in a more general sense? We leave these questions for future work.

Acknowledgments

R.F.B. was partially supported by the National Science Foundation via grants DMS-1654076 and DMS-2023109, and by the Office of Naval Research via grant N00014-20-1-2337.

Appendix A: Additional proofs

A.1. Proof of Lemma 1

We first verify that $\widetilde{\mathcal{L}}^\rho$ is β-Lipschitz, where $\beta = L\sqrt{\frac{a_1}{a_0}}$. For any $w \neq w' \in \mathbb{R}^d$ we have

$$\begin{aligned}
\left|\widetilde{\mathcal{L}}^\rho(w) - \widetilde{\mathcal{L}}^\rho(w')\right| &= \left|\mathbb{E}[\ell(X^\top w \cdot \widetilde{Y}) - \ell(X^\top w' \cdot \widetilde{Y})]\right| \\
&\leqslant \mathbb{E}\left[\left|\ell(X^\top w \cdot \widetilde{Y}) - \ell(X^\top w' \cdot \widetilde{Y})\right|\right] \\
&\leqslant \mathbb{E}\left[L \cdot \left|X^\top w \cdot \widetilde{Y} - X^\top w' \cdot \widetilde{Y}\right|\right] \quad \text{since } \ell \text{ is } L\text{-Lipschitz by Assumption 1} \\
&= L\mathbb{E}\left[\left|X^\top (w - w')\right|\right] \quad \text{since } \widetilde{Y} \in \{\pm 1\} \\
&= L\|w - w'\| \cdot \mathbb{E}[|X^\top u|] \quad \text{where } u = \frac{w - w'}{\|w - w'\|} \\
&\leqslant \beta \cdot \|w - w'\|,
\end{aligned}$$

where the last inequality follows from Assumption 2 via the calculation

$$a_1 \geqslant \mathbb{E}\left[e^{a_0 |X^\top v|^2}\right] \geqslant a_0 \cdot \mathbb{E}[|X^\top v|^2] \geqslant a_0 \cdot \mathbb{E}[|X^\top v|]^2. \tag{3}$$

We therefore have that $\widetilde{\mathcal{L}}^\rho$ is β-Lipschitz. Note that the above argument also holds for $\rho = 0$, implying that $\mathcal{L}$ is also β-Lipschitz.

Now fix $t = C_0 \rho^{-1/2}$ for any $C_0 > \sqrt{\frac{8c_1 a_2^2}{c_2 \gamma \log 2}}$. We will show that, for any $u \in \mathbb{S}^{d-1}$,

$$\widetilde{\mathcal{L}}^\rho(t \cdot u) > \widetilde{\mathcal{L}}^\rho(0.5t \cdot u).$$

First we calculate

$$\mathbb{E}\left[|X^\top u| \cdot \mathbb{1}\left\{X^\top u \cdot \widetilde{Y} < 0\right\}\right] \geqslant \rho \cdot \mathbb{E}[|X^\top u|] \geqslant \rho \cdot \frac{\log 2}{2a_2}$$

where the first inequality holds by definition of the distribution of the corrupted label $\widetilde{Y}$ (since $\mathbb{P}\{\widetilde{Y} = +1 \mid X\} \in [\rho, 1-\rho]$ holds almost surely), while for the second inequality, by Jensen's inequality together with Assumption 2,

$$e^{-2a_2\mathbb{E}[|X^\top u|]} \leqslant \mathbb{E}[e^{-2a_2|X^\top u|}] \leqslant \frac{a_2}{2a_2} = \frac{1}{2},$$

so

$$\mathbb{E}[|X^\top u|] \geqslant \frac{\log 2}{2a_2}. \tag{4}$$

We also know that

$$\ell(-t \cdot |X^\top u|) - \ell(-0.5t \cdot |X^\top u|) \geqslant \gamma \cdot 0.5t \cdot |X^\top u|,$$

by Assumption 1, and so

$$\mathbb{E}\left[\left(\ell(t \cdot X^\top u \cdot \widetilde{Y}) - \ell(0.5t \cdot X^\top u \cdot \widetilde{Y})\right) \cdot \mathbb{1}\left\{X^\top u \cdot \widetilde{Y} < 0\right\}\right] \\ \geqslant \mathbb{E}\left[\gamma \cdot 0.5t \cdot |X^\top u| \cdot \mathbb{1}\left\{X^\top u \cdot \widetilde{Y} < 0\right\}\right] \geqslant \gamma \cdot 0.5t \cdot \rho \cdot \frac{\log 2}{2a_2}.$$

We therefore have

$$\begin{aligned}
&\widetilde{\mathcal{L}}^\rho(t \cdot u) - \widetilde{\mathcal{L}}^\rho(0.5t \cdot u) \\
&= \mathbb{E}\left[\ell(t \cdot X^\top u \cdot \widetilde{Y}) - \ell(0.5t \cdot X^\top u \cdot \widetilde{Y})\right] \\
&= \mathbb{E}\left[\left(\ell(t \cdot X^\top u \cdot \widetilde{Y}) - \ell(0.5t \cdot X^\top u \cdot \widetilde{Y})\right) \cdot \mathbb{1}\left\{X^\top u \cdot \widetilde{Y} < 0\right\}\right] \\
&\qquad + \mathbb{E}\left[\left(\ell(t \cdot X^\top u \cdot \widetilde{Y}) - \ell(0.5t \cdot X^\top u \cdot \widetilde{Y})\right) \cdot \mathbb{1}\left\{X^\top u \cdot \widetilde{Y} \geqslant 0\right\}\right] \\
&\geqslant \gamma \cdot 0.5t \cdot \rho \cdot \frac{\log 2}{2a_2} + \mathbb{E}\left[\left(\ell(t \cdot |X^\top u|) - \ell(0.5t \cdot |X^\top u|)\right) \cdot \mathbb{1}\left\{X^\top u \cdot \widetilde{Y} \geqslant 0\right\}\right] \\
&\geqslant \gamma \cdot 0.5t \cdot \rho \cdot \frac{\log 2}{2a_2} - \mathbb{E}\left[\ell(0.5t \cdot |X^\top u|)\right] \\
&\geqslant \gamma \cdot 0.5t \cdot \rho \cdot \frac{\log 2}{2a_2} - c_1\mathbb{E}\left[e^{-c_2 \cdot 0.5t \cdot |X^\top u|}\right] \text{ by Assumption 1} \\
&\geqslant \gamma \cdot 0.5t \cdot \rho \cdot \frac{\log 2}{2a_2} - \frac{c_1 a_2}{c_2 \cdot 0.5t} \text{ by Assumption 2} \\
&> 0 \text{ by definition of } t.
\end{aligned}$$

In particular, this implies that $\widetilde{\mathcal{L}}^\rho(tu) > \inf_{w \in \mathbb{R}^d} \widetilde{\mathcal{L}}^\rho(w)$ for all $u \in \mathbb{S}^{d-1}$. Since $w \mapsto \widetilde{\mathcal{L}}^\rho(w)$ is continuous as shown above, this implies that $\widetilde{\mathcal{L}}^\rho(w)$ attains its infimum, and any $\widetilde{w}^\rho_* \in \operatorname{argmin}_{w \in \mathbb{R}^d} \widetilde{\mathcal{L}}^\rho(w)$ must satisfy $\|\widetilde{w}^\rho_*\| \leqslant t$.

Next we bound $\mathcal{L}(\widetilde{w}^\rho_*)$ for any $\widetilde{w}^\rho_* \in \operatorname{argmin}_{w \in \mathbb{R}^d} \widetilde{\mathcal{L}}^\rho(w)$. First note that the corrupted risk can be written as

$$\widetilde{\mathcal{L}}^\rho(w) = (1-2\rho) \cdot \mathcal{L}(w) + 2\rho \cdot \mathsf{R}(w) = (1-\rho)\mathcal{L}(w) + \rho\mathcal{L}(-w). \tag{5}$$

Applying (5) with $w = \widetilde{w}^\rho_*$ we obtain

$$\widetilde{\mathcal{L}}^\rho(\widetilde{w}^\rho_*) = (1-\rho)\mathcal{L}(\widetilde{w}^\rho_*) + \rho\mathcal{L}(-\widetilde{w}^\rho_*),$$

and similarly applying (5) with $w = -\widetilde{w}^\rho_*$ we obtain

$$\widetilde{\mathcal{L}}^\rho(-\widetilde{w}^\rho_*) = (1-\rho)\mathcal{L}(-\widetilde{w}^\rho_*) + \rho\mathcal{L}(\widetilde{w}^\rho_*).$$

Since $\widetilde{\mathcal{L}}^\rho(\widetilde{w}_*^\rho) \leqslant \widetilde{\mathcal{L}}^\rho(-\widetilde{w}_*^\rho)$ by optimality of $\widetilde{w}_*^\rho$, and $\rho < \frac{1}{2}$ by assumption, this proves that $\mathcal{L}(\widetilde{w}_*^\rho) \leqslant \mathcal{L}(-\widetilde{w}_*^\rho)$ and therefore,

$$\mathcal{L}(\widetilde{w}_*^\rho) \leqslant \widetilde{\mathcal{L}}^\rho(\widetilde{w}_*^\rho).$$

Next, fix any $w \in \mathbb{R}^d$. First consider the case that $\|w\| \leqslant c\rho^{-1/2}$, where $c = \sqrt{\frac{c_1 a_2}{2\beta c_2}}$. Then

$$\begin{aligned}
\widetilde{\mathcal{L}}^\rho(\widetilde{w}_*^\rho) - \mathcal{L}(w) &\leqslant \widetilde{\mathcal{L}}^\rho(w) - \mathcal{L}(w) \quad \text{by optimality of } \widetilde{w}_*^\rho \\
&= \rho\left(\mathcal{L}(-w) - \mathcal{L}(w)\right) \quad \text{by (5)} \\
&\leqslant 2\rho\beta \cdot c\rho^{-1/2} \\
&= 2\beta c \rho^{1/2},
\end{aligned}$$

where the last inequality holds since $\mathcal{L}$ is β-Lipschitz.

Next consider the case that $\|w\| > c\rho^{-1/2}$. Let $u = w/\|w\|$ and $t = c\rho^{-1/2}$. Then by the reasoning above, we have

$$\widetilde{\mathcal{L}}^\rho(\widetilde{w}_*^\rho) - \mathcal{L}(tu) \leqslant 2\beta c \rho^{1/2}.$$

Next, let $Z_u = X^\top u \cdot Y$, then we have

$$\begin{aligned}
&\mathcal{L}(tu) - \mathcal{L}(w) = \mathbb{E}[\ell(t \cdot Z_u) - \ell(\|w\| \cdot Z_u)] \\
&= \mathbb{E}[(\ell(t \cdot Z_u) - \ell(\|w\| \cdot Z_u)) \cdot \mathbb{1}\{Z_u > 0\}] + \mathbb{E}[(\ell(t \cdot Z_u) - \ell(\|w\| \cdot Z_u)) \cdot \mathbb{1}\{Z_u < 0\}] \\
&\leqslant \mathbb{E}[(\ell(t \cdot Z_u) - \ell(\|w\| \cdot Z_u)) \cdot \mathbb{1}\{Z_u > 0\}] \quad \text{since } \|w\| > t \text{ and } \ell \text{ is nonincreasing} \\
&\leqslant \mathbb{E}[\ell(t \cdot Z_u) \cdot \mathbb{1}\{Z_u > 0\}] \quad \text{since } \ell \text{ is nonnegative} \\
&\leqslant c_1 \mathbb{E}[e^{-c_2 t |X^\top u|}] \quad \text{by Assumption 1} \\
&\leqslant c_1 \cdot \frac{a_2}{c_2 t} \quad \text{by Assumption 2} \\
&= \frac{c_1 a_2}{c_2 c} \cdot \rho^{1/2}.
\end{aligned}$$

Therefore, for this second case, we have shown that

$$\widetilde{\mathcal{L}}^\rho(\widetilde{w}_*^\rho) - \mathcal{L}(w) \leqslant \left(2\beta c + \frac{c_1 a_2}{c_2 c}\right) \cdot \rho^{1/2} = \sqrt{\frac{8\beta c_1 a_2}{c_2}} \cdot \rho^{1/2}.$$

Combining the two cases, we have shown that

$$\mathcal{L}(\widetilde{w}_*^\rho) \leqslant \widetilde{\mathcal{L}}^\rho(\widetilde{w}_*^\rho) \leqslant \mathcal{L}(w) + \sqrt{\frac{8\beta c_1 a_2}{c_2}} \cdot \rho^{1/2}$$

for all $w \in \mathbb{R}^d$, which proves the desired inequality with

$$C_1 = \sqrt{\frac{8\beta c_1 a_2}{c_2}}.$$

Now we turn to the corrupted estimator $\widetilde{w}_n^\rho$. First we will need a lemma to establish some concentration results.

Lemma 3. *Suppose Assumptions 1 and 2 hold. Fix any $\alpha > 0$, $\rho \in (0, \frac{1}{2})$, $t > 0$, and $r > 0$. Then with probability at least $1 - n^{-\alpha}$, it holds that*

$$\inf_{u \in \mathbb{S}^{d-1}} \left\{ \frac{1}{n} \sum_{i=1}^n \max\left\{0, -X_i^\top u \cdot \tilde{Y}_i\right\} \right\} \geqslant r_1 \rho - r_2 \cdot \frac{d \log n}{n} \tag{6}$$

and

$$\sup_{u \in \mathbb{S}^{d-1}} \left\{ \frac{1}{n} \sum_{i=1}^n e^{-t|X_i^\top u|} \right\} \leqslant \frac{r_3}{t} + r_4 \sqrt{\frac{d \log n}{n}} \tag{7}$$

and

$$\sup_{\|w\|\leqslant r}\left|\widetilde{\mathcal{L}}_n^\rho(w)-\widetilde{\mathcal{L}}^\rho(w)\right|\leqslant r_5\cdot r\cdot\sqrt{\frac{d\log n}{n}},\tag{8}$$

where $r_1, r_2, r_3, r_4, r_5 > 0$ depend only on α and on the constants in Assumptions 1 and 2, and not on n, d, r, or t.

We are now ready to prove the remainder of Lemma 1. First we bound $\|\widetilde{w}_n^\rho\|$. Define $C = \frac{2r_2}{r_1}$ and fix $t = C_0\rho^{-1/2}$ for any $C_0 > \max\left\{2\sqrt{\frac{4c_1\left(2c_2^{-1}r_3\right)}{\gamma r_1}}, \frac{8c_1\left(C^{-1/2}r_4\right)}{\gamma r_1}\right\}$, which therefore satisfies

$$C_0 > \sqrt{\frac{4c_1\left(2c_2^{-1}r_3 + C_0C^{-1/2}r_4\right)}{\gamma r_1}}.$$

We will show that, for any $u \in \mathbb{S}^{d-1}$,

$$\widetilde{\mathcal{L}}_n^\rho(t\cdot u) > \widetilde{\mathcal{L}}_n^\rho(0.5t\cdot u).$$

Then assuming $\rho \geqslant C\cdot\frac{d\log n}{n}$, the bound (6) in Lemma 3 implies that

$$\frac{1}{n}\sum_{i=1}^n |X_i^\top u|\cdot\mathbb{1}\left\{X_i^\top u\cdot\widetilde{Y}_i<0\right\} = \frac{1}{n}\sum_{i=1}^n \max\left\{0, -X_i^\top u\cdot\tilde{Y}_i\right\}\geqslant\frac{r_1}{2}\cdot\rho,$$

for all $u \in \mathbb{S}^{d-1}$. Furthermore, since $t = C_0\rho^{-1/2}$, the bound (7) in Lemma 3 (applied with $0.5c_2t$ in place of t) together with our assumption $\rho \geqslant C\cdot\frac{d\log n}{n}$ implies that

$$\frac{1}{n}\sum_{i=1}^n e^{-c_2\cdot 0.5t|X_i^\top u|}\leqslant\frac{2c_2^{-1}r_3+C_0C^{-1/2}r_4}{t}$$

for all $u \in \mathbb{S}^{d-1}$. Following identical arguments as in the population case, we have

$$\widetilde{\mathcal{L}}_n^\rho(t\cdot u)-\widetilde{\mathcal{L}}_n^\rho(0.5t\cdot u)\geqslant\gamma\cdot 0.5t\cdot\rho\cdot r_1/2 - c_1\cdot\frac{2c_2^{-1}r_3+C_0C^{-1/2}r_4}{t}>0$$

for all $u \in \mathbb{S}^{d-1}$, where the last step holds by definition of t and of C_0. Since $\widetilde{\mathcal{L}}_n^\rho$ is continuous (because we have assumed the loss ℓ is continuous), as for the population case this again proves that $\widetilde{\mathcal{L}}_n^\rho(w)$ must attain its infimum, and that any $w \in \operatorname{argmin}_{w\in\mathbb{R}^d}\widetilde{\mathcal{L}}_n^\rho(w)$ must satisfy $\|w\| \leqslant t$.

Finally, the bound $\sup_{\|w\|\leqslant C_0\rho^{-1/2}}\left|\widetilde{\mathcal{L}}_n^\rho(w)-\widetilde{\mathcal{L}}^\rho(w)\right| \leqslant C_2\rho^{-1/2}\sqrt{\frac{d\log n}{n}}$ follows immediately from the bound (8) in Lemma 3, by setting $C_2 = C_0r_5$.

A.2. Proof of Lemma 3

First, we prove (6). The distribution of $(X, \widetilde{Y})$ can equivalently be represented as

$$(X,\widetilde{Y}) = \big(X, (1-R)\cdot Y + R\cdot Z\big),$$

where $R \sim \text{Bernoulli}(2\rho)$ is generated independently from (X, Y), and $Z \sim \text{Unif}\{\pm 1\}$ is generated independently from (X, Y, R). Let (X_i, Y_i, R_i, Z_i) generate the n i.i.d. data points. Furthermore, define

$$\bar{X} = X\cdot\min\left\{1, \frac{4\mathbb{E}[\|X\|]}{\|X\|}\right\}.$$

and

$$\bar{X}_i = X_i\cdot\min\left\{1, \frac{4\mathbb{E}[\|X\|]}{\|X_i\|}\right\}.$$

Then we can check that, for all $u \in \mathbb{S}^{d-1}$,

$$\frac{1}{n}\sum_{i=1}^{n} \max\left\{0, -X_i^\top u \cdot \tilde{Y}_i\right\} \geqslant \frac{1}{n}\sum_{i=1}^{n} \max\left\{0, -\bar{X}_i^\top u \cdot \tilde{Y}_i\right\} \geqslant \frac{1}{n}\sum_{i=1}^{n} \max\left\{0, -\bar{X}_i^\top u \cdot R_i \cdot Z_i\right\}.$$

Define

$$\Delta = \sup_{u\in\mathbb{S}^{d-1}} \left| \frac{1}{n}\sum_{i=1}^{n} \max\left\{0, -\bar{X}_i^\top u \cdot R_i \cdot Z_i\right\} - \mathbb{E}[\max\left\{0, -\bar{X}^\top u \cdot R \cdot Z\right\}] \right|.$$

We can verify that, since $\bar{X}, R, Z$ are independent, by definition of their distributions we have

$$\mathbb{E}\left[\max\left\{0, -\bar{X}^\top u \cdot R \cdot Z\right\}\right] \geqslant \rho \cdot \mathbb{E}\left[|\bar{X}^\top u|\right].$$

Furthermore, by Jensen's inequality,

$$\begin{aligned}\exp\left\{-4a_2\mathbb{E}\left[|\bar{X}^\top u|\right]\right\} \leqslant \mathbb{E}\left[e^{-4a_2|\bar{X}^\top u|}\right] &\leqslant \mathbb{E}\left[e^{-4a_2|X^\top u|}\right] + \mathbb{P}\{\|X\| > 4\mathbb{E}[\|X\|]\} \\ &\leqslant \frac{a_2}{4a_2} + \frac{\mathbb{E}[\|X\|]}{4\mathbb{E}[\|X\|]} = \frac{1}{2},\end{aligned}$$

where the last inequality applies Assumption 2 together with Markov's inequality. Rearranging terms, then,

$$\mathbb{E}\left[|\bar{X}^\top u|\right] \geqslant \frac{\log 2}{4a_2}.$$

Therefore, combining everything we have shown so far, it holds deterministically that

$$\inf_{u\in\mathbb{S}^{d-1}} \left\{ \frac{1}{n}\sum_{i=1}^{n} \max\left\{0, -X_i^\top u \cdot \tilde{Y}_i\right\} \right\} \geqslant \rho \cdot \frac{\log 2}{4a_2} - \Delta.$$

Now we need to bound Δ with high probability.

By the symmetrization inequality Koltchinskii [10, Theorem 2.1] we have

$$\mathbb{E}[\Delta] \leqslant 2\mathbb{E}\left[\sup_{u\in\mathbb{S}^{d-1}} \left| \frac{1}{n}\sum_{i=1}^{n} \xi_i \cdot \max\left\{0, -\bar{X}_i^\top u \cdot R_i \cdot Z_i\right\} \right|\right],$$

where the last expectation is taken with respect to the i.i.d. data $(\bar{X}_i, \tilde{Y}_i)$ as well as i.i.d. Rademacher random variables $\xi_i \overset{\text{iid}}{\sim} \text{Unif}\{\pm 1\}$. Since $t \mapsto \max\{0, -t\}$ is 1-Lipschitz, the contraction inequality Koltchinskii [10, Theorem 2.2] verifies that

$$\mathbb{E}[\Delta] \leqslant 4\mathbb{E}\left[\sup_{u\in\mathbb{S}^{d-1}} \left| \frac{1}{n}\sum_{i=1}^{n} \xi_i \cdot \bar{X}_i^\top u \cdot R_i \cdot Z_i \right|\right].$$

Furthermore, deterministically we have

$$\left| \frac{1}{n}\sum_{i=1}^{n} \xi_i \cdot \bar{X}_i^\top u \cdot R_i \cdot Z_i \right| = \left| u^\top \left(\frac{1}{n}\sum_{i=1}^{n} \xi_i \cdot R_i \cdot Z_i \cdot \bar{X}_i \right) \right| \leqslant \left\| \frac{1}{n}\sum_{i=1}^{n} \xi_i \cdot R_i \cdot Z_i \cdot \bar{X}_i \right\|,$$

and so combining everything so far, we have shown that

$$\mathbb{E}[\Delta] \leqslant 4\mathbb{E}\left[\left\| \frac{1}{n}\sum_{i=1}^{n} \xi_i \cdot R_i \cdot Z_i \cdot \bar{X}_i \right\|\right].$$

Moreover, we can see that $(\bar{X}_i, \xi_i \cdot Z_i)$ is equal in distribution to $(\bar{X}_i, \xi_i)$ (since $Z_i \in \{\pm 1\}$ while $\xi_i \sim \text{Unif}\{\pm 1\}$ is drawn independently from the data), and so

$$\mathbb{E}[\Delta] \leqslant 4\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \xi_i \cdot \bar{X}_i \cdot R_i\right\|\right].$$

Finally,

$$\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \xi_i \cdot \bar{X}_i \cdot R_i\right\|\right]^2 \leqslant \mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \xi_i \cdot \bar{X}_i \cdot R_i\right\|^2\right] = \frac{1}{n^2}\sum_{j=1}^d \mathbb{E}\left[\left(\sum_{i=1}^n \bar{X}_{ij} R_i \xi_i\right)^2\right]$$
$$= \frac{1}{n^2}\sum_{j=1}^d \sum_{i=1}^n \mathbb{E}[\bar{X}_{ij}^2 R_i^2] = \frac{1}{n^2}\sum_{i=1}^n 2\rho \mathbb{E}[\|\bar{X}_i\|^2] \leqslant \frac{1}{n} \cdot 16\mathbb{E}[\|X\|]^2 \cdot 2\rho,$$

since by definition, it holds deterministically that $\|\bar{X}_i\| \leqslant 4\mathbb{E}[\|X\|]$, while $R_i \sim \text{Bernoulli}(2\rho)$ is independent from X_i. Combining everything so far,

$$\mathbb{E}[\Delta] \leqslant 4\sqrt{\frac{1}{n} \cdot 16\mathbb{E}[\|X\|]^2 \cdot 2\rho}.$$

Next, since for all $u \in \mathbb{S}^{d-1}$ we have

$$\mathbb{E}[\max\left\{0, -\bar{X}^\top u \cdot R \cdot Z\right\}^2] \leqslant 2\rho \cdot \left(4\mathbb{E}[\|X\|]\right)^2$$

and

$$0 \leqslant \max\left\{0, -\bar{X}^\top u \cdot R \cdot Z\right\} \leqslant 4\mathbb{E}[\|X\|] \text{ almost surely,}$$

applying Koltchinskii [10, Bousquet bound, Section 2.3] yields the concentration result

$$\mathbb{P}\left\{\Delta \leqslant \mathbb{E}[\Delta] + \sqrt{\frac{2\log(3n^\alpha) \cdot (2\rho \cdot 16\mathbb{E}[\|X\|]^2 + 4\mathbb{E}[\|X\|] \cdot 2\mathbb{E}[\Delta])}{n}} + 4\mathbb{E}[\|X\|] \cdot \frac{\log(3n^\alpha)}{3n}\right\}$$
$$\geqslant 1 - \frac{1}{3n^\alpha}.$$

Furthermore, Assumption 2 together with Jensen's inequality implies

$$e^{a_0 \mathbb{E}[\|X\|^2]/d} \leqslant e^{a_0 \max_{1\leqslant j\leqslant d} \mathbb{E}[|X_j|^2]} \leqslant \max_{1\leqslant j\leqslant d} \mathbb{E}[e^{a_0 |X_j|^2}] \leqslant a_1$$

and so $\mathbb{E}[\|X\|] \leqslant \mathbb{E}[\|X\|^2]^{1/2} \leqslant \sqrt{\frac{d\log a_1}{a_0}}$. Combined with our bound on $\mathbb{E}[\Delta]$, we can verify that this bound can be relaxed to

$$\mathbb{P}\left\{\Delta \leqslant r'\left(\sqrt{\rho \cdot \frac{d\log n}{n}} + \frac{d\log n}{n}\right)\right\} \geqslant 1 - \frac{1}{3n^\alpha}$$

where r' is chosen appropriately as a function of α, a_0, and a_1. Therefore, we have shown that with probability at least $1 - \frac{1}{3n^\alpha}$,

$$\inf_{u\in\mathbb{S}^{d-1}}\left\{\frac{1}{n}\sum_{i=1}^n \max\left\{0, -X_i^\top u \cdot \tilde{Y}_i\right\}\right\} \geqslant \rho \cdot \frac{\log 2}{4a_2} - r'\left(\sqrt{\rho \cdot \frac{d\log n}{n}} + \frac{d\log n}{n}\right),$$

which is sufficient to verify (6) with r_1, r_2 chosen appropriately, since it holds that $\sqrt{\rho \cdot \frac{d\log n}{n}} \leqslant \frac{r''\rho}{2} + \frac{d\log n}{2r''n}$ for all $r'' > 0$.

Next we prove (7). Note that, comparing the two terms in the desired upper bound and noting that $1/t$ is only dominant if $t \leqslant \sqrt{\frac{n}{d \log n}}$, we can see that it suffices to prove the result for $t \leqslant \sqrt{\frac{n}{d \log n}}$, since $t \mapsto \sup_{u \in \mathbb{S}^{d-1}} \left\{ \frac{1}{n} \sum_{i=1}^n e^{-t|X_i^\top u|} \right\}$ is monotone nonincreasing in t.

We have

$$\sup_{u \in \mathbb{S}^{d-1}} \left\{ \frac{1}{n} \sum_{i=1}^n e^{-t|X_i^\top u|} \right\} \leqslant \sup_{u \in \mathbb{S}^{d-1}} \left\{ \frac{1}{n} \sum_{i=1}^n e^{-t|\bar{X}_i^\top u|} \right\},$$

where, changing the definition of $\bar{X}$ and $\bar{X}_i$, we let

$$\bar{X} = X \cdot \min \left\{ 1, \frac{t\mathbb{E}[\|X\|]}{\|X\|} \right\}.$$

and analogously

$$\bar{X}_i = X_i \cdot \min \left\{ 1, \frac{t\mathbb{E}[\|X\|]}{\|X_i\|} \right\}.$$

Next fix $\epsilon > 0$, and take a covering $u_1, \ldots, u_M$ of $\mathbb{S}^{d-1}$ such that

$$\sup_{u \in \mathbb{S}^{d-1}} \left\{ \min_{m=1,\ldots,M} \|u - u_m\| \right\} \leqslant \epsilon.$$

By Lorentz, Golitschek and Makovoz [15, Chapter 15], for any $\epsilon > 0$ we can construct a set with this property of size $M \leqslant (3/\epsilon)^d$. Then for any $u \in \mathbb{S}^{d-1}$, if we find m such that $\|u - u_m\| \leqslant \epsilon$, we have

$$e^{-t|\bar{X}_i^\top u|} \leqslant e^{-t|\bar{X}_i^\top u_m|} + t\|\bar{X}_i\| \cdot \epsilon \leqslant e^{-t|\bar{X}_i^\top u_m|} + t^2 \mathbb{E}[\|X\|] \cdot \epsilon,$$

since $e^{-t|x|}$ is t-Lipschitz over $x \in \mathbb{R}$. Therefore,

$$\sup_{u \in \mathbb{S}^{d-1}} \left\{ \frac{1}{n} \sum_{i=1}^n e^{-t|X_i^\top u|} \right\} \leqslant t^2 \mathbb{E}[\|X\|] \cdot \epsilon + \max_{m=1,\ldots,M} \left\{ \frac{1}{n} \sum_{i=1}^n e^{-t|\bar{X}_i^\top u_m|} \right\}.$$

Next, for each m, by Hoeffding's inequality,

$$\mathbb{P} \left\{ \frac{1}{n} \sum_{i=1}^n e^{-t|\bar{X}_i^\top u_m|} - \mathbb{E}[e^{-t|\bar{X}^\top u_m|}] > \sqrt{\frac{\log(3Mn^\alpha)}{2n}} \right\} \leqslant \frac{1}{3Mn^\alpha}.$$

Furthermore,

$$\mathbb{E}[e^{-t|\bar{X}^\top u_m|}] \leqslant \mathbb{E}[e^{-t|X^\top u_m|}] + \mathbb{P}\{\|X\| > t\mathbb{E}[\|X\|]\} \leqslant \frac{a_2 + 1}{t},$$

by applying Assumption 2 together with Markov's inequality. Therefore, combining everything, with probability at least $1 - \frac{1}{3n^\alpha}$,

$$\sup_{u \in \mathbb{S}^{d-1}} \left\{ \frac{1}{n} \sum_{i=1}^n e^{-t|X_i^\top u|} \right\} \leqslant t^2 \mathbb{E}[\|X\|] \cdot \epsilon + \sqrt{\frac{\log(3 \cdot (3/\epsilon)^d \cdot n^\alpha)}{2n}} + \frac{a_2 + 1}{t}.$$

Since we have assumed that $t \leqslant n$, taking $\epsilon = n^{-2.5}$ we obtain

$$\sup_{u \in \mathbb{S}^{d-1}} \left\{ \frac{1}{n} \sum_{i=1}^n e^{-t|X_i^\top u|} \right\} \leqslant \frac{\mathbb{E}[\|X\|]}{\sqrt{n}} + \sqrt{\frac{\log(3 \cdot (3n^{2.5})^d \cdot n^\alpha)}{2n}} + \frac{a_2 + 1}{t},$$

which clearly satisfies (7) with r_3, r_4 chosen appropriately, since as shown before, $\mathbb{E}[\|X\|] \leqslant \sqrt{\frac{d \log a_1}{a_0}}$.

Finally we prove (8). We first bound the quantity in the expected value. We have

$$\mathbb{E}\left[\sup_{\|w\|\leqslant r}\left|\widetilde{\mathcal{L}}_n^\rho(w)-\widetilde{\mathcal{L}}^\rho(w)\right|\right]=\mathbb{E}\left[\sup_{\|w\|\leqslant r}\left|\frac{1}{n}\sum_{i=1}^n\left(\ell(X_i^\top w\cdot\widetilde{Y}_i)-\mathbb{E}[\ell(X_i^\top w\cdot\widetilde{Y}_i)]\right)\right|\right]$$
$$\leqslant 2\mathbb{E}\left[\sup_{\|w\|\leqslant r}\left|\frac{1}{n}\sum_{i=1}^n\xi_i\ell(X_i^\top w\cdot\widetilde{Y}_i)\right|\right],$$

by the symmetrization inequality Koltchinskii [10, Theorem 2.1], where the last expectation is taken with respect to the i.i.d. data $(\bar{X}_i,\widetilde{Y}_i)$ as well as i.i.d. Rademacher random variables $\xi_i\overset{\text{iid}}{\sim}\text{Unif}\{\pm1\}$. Next, the contraction inequality Koltchinskii [10, Theorem 2.2] verifies that

$$\mathbb{E}\left[\sup_{\|w\|\leqslant r}\left|\frac{1}{n}\sum_{i=1}^n\xi_i\ell(X_i^\top w\cdot\widetilde{Y}_i)\right|\right]\leqslant 2L\mathbb{E}\left[\sup_{\|w\|\leqslant r}\left|\frac{1}{n}\sum_{i=1}^n\xi_i\cdot X_i^\top w\cdot\widetilde{Y}_i\right|\right],$$

since ℓ is L-Lipschitz by Assumption 1. Furthermore, deterministically we have

$$\left|\frac{1}{n}\sum_{i=1}^n\xi_i\cdot X_i^\top w\cdot\widetilde{Y}_i\right|=\left|w^\top\left(\frac{1}{n}\sum_{i=1}^n\xi_i\cdot\widetilde{Y}_i\cdot X_i\right)\right|\leqslant\|w\|\cdot\left\|\frac{1}{n}\sum_{i=1}^n\xi_i\cdot\widetilde{Y}_i\cdot X_i\right\|,$$

and so combining everything so far, we have shown that

$$\mathbb{E}\left[\sup_{\|w\|\leqslant r}\left|\widetilde{\mathcal{L}}_n^\rho(w)-\widetilde{\mathcal{L}}^\rho(w)\right|\right]\leqslant 4Lr\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n\xi_i\cdot\widetilde{Y}_i\cdot X_i\right\|\right].$$

Moreover, we can see that $(X_i,\xi_i\cdot\widetilde{Y}_i)$ is equal in distribution to (X_i,ξ_i) (since $\widetilde{Y}_i\in\{\pm1\}$ while $\xi_i\sim\text{Unif}\{\pm1\}$ is drawn independently from $(X_i,\widetilde{Y}_i)$), and so

$$\mathbb{E}\left[\sup_{\|w\|\leqslant r}\left|\widetilde{\mathcal{L}}_n^\rho(w)-\widetilde{\mathcal{L}}^\rho(w)\right|\right]\leqslant 4Lr\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n\xi_i\cdot X_i\right\|\right].$$

Finally,

$$\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n\xi_i\cdot X_i\right\|\right]^2\leqslant\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n\xi_i\cdot X_i\right\|^2\right]=\frac{1}{n^2}\sum_{j=1}^d\mathbb{E}\left[\left(\sum_{i=1}^n X_{ij}\xi_i\right)^2\right]$$
$$=\frac{1}{n^2}\sum_{j=1}^d\sum_{i=1}^n\mathbb{E}[X_{ij}^2]=\frac{1}{n}\mathbb{E}[\|X\|^2]\leqslant\frac{d}{n}\cdot\frac{\log a_1}{a_0},$$

since $\mathbb{E}[\|X\|^2]\leqslant\frac{d\log a_1}{a_0}$ as calculated above. Therefore,

$$\mathbb{E}\left[\sup_{\|w\|\leqslant r}\left|\widetilde{\mathcal{L}}_n^\rho(w)-\widetilde{\mathcal{L}}^\rho(w)\right|\right]\leqslant\frac{4Lr\sqrt{\log a_1}}{\sqrt{a_0}}\cdot\sqrt{\frac{d}{n}}.$$

Next we prove that the quantity $\sup_{\|w\|\leqslant r}\left|\widetilde{\mathcal{L}}_n^\rho(w)-\widetilde{\mathcal{L}}^\rho(w)\right|$ concentrates around its expectation. First, let $(X',\widetilde{Y}')$ be an i.i.d. draw from the distribution of $(X,\widetilde{Y})$. For $\lambda\geqslant 0$, we calculate

$$\mathbb{E}\left[\frac{1}{2}e^{\lambda\|X\widetilde{Y}-X'\widetilde{Y}'\|}+\frac{1}{2}e^{-\lambda\|X\widetilde{Y}-X'\widetilde{Y}'\|}\right]\leqslant\mathbb{E}\left[e^{\lambda^2\|X\widetilde{Y}-X'\widetilde{Y}'\|^2/2}\right]$$
$$\leqslant\mathbb{E}\left[e^{\lambda^2\cdot(\|X\widetilde{Y}\|^2+\|X'\widetilde{Y}'\|^2)}\right]=\mathbb{E}\left[e^{\lambda^2\cdot\|X\widetilde{Y}\|^2}\right]^2=\mathbb{E}\left[e^{\lambda^2\cdot\|X\|^2}\right]^2$$
$$=\mathbb{E}\left[e^{\lambda^2\cdot\sum_{j=1}^d|X_j|^2}\right]^2\leqslant\mathbb{E}\left[\frac{1}{d}\sum_{j=1}^d e^{d\lambda^2\cdot|X_j|^2}\right]^2,$$

by the AM–GM inequality. Applying Assumption 2, we then obtain

$$\mathbb{E}\left[\frac{1}{2}e^{\lambda\|X\tilde{Y}-X'\tilde{Y}'\|}+\frac{1}{2}e^{-\lambda\|X\tilde{Y}-X'\tilde{Y}'\|}\right]\leqslant a_1^{\frac{2\lambda^2 d}{a_0}}$$

as long as $\lambda^2 \leqslant a_0/d$. Following the proof of Kontorovich [11, Theorem 1], since $\sup_{\|w\|\leqslant r}\left|\widetilde{\mathcal{L}}_n^\rho(w)-\widetilde{\mathcal{L}}^\rho(w)\right|$ is a $\frac{Lr}{n}$-Lipschitz function of each data point product $X_i\cdot\widetilde{Y}_i$,

$$\mathbb{P}\left\{\sup_{\|w\|\leqslant r}\left|\widetilde{\mathcal{L}}_n^\rho(w)-\widetilde{\mathcal{L}}^\rho(w)\right|-\mathbb{E}\left[\sup_{\|w\|\leqslant r}\left|\widetilde{\mathcal{L}}_n^\rho(w)-\widetilde{\mathcal{L}}^\rho(w)\right|\right]>\frac{Lr}{n}\cdot\sqrt{\frac{8nd\log a_1\cdot\log(3n^\alpha)}{a_0}}\right\}$$
$$\leqslant\exp\left\{\frac{2n\lambda^2 d\log a_1}{a_0}-\lambda\cdot\sqrt{\frac{8nd\log a_1\cdot\log(3n^\alpha)}{a_0}}\right\}.$$

Taking

$$\lambda=\frac{a_0}{4nd\log a_1}\cdot\sqrt{\frac{8nd\log a_1\cdot\log(3n^\alpha)}{a_0}}$$

(which clearly satisfies $\lambda\leqslant\sqrt{\frac{a_0}{d}}$ for sufficiently large n), this probability is bounded by $\frac{1}{3n^\alpha}$. (If instead n is not sufficiently large (i.e., $\lambda>\sqrt{\frac{a_0}{d}}$), then the guarantee (8) holds trivially.) Combining everything, and choosing r_5 appropriately, we have established (8).

References

[1] Blanchard, G., Flaska, M., Handy, G., Pozzi, S., Scott, C. et al. (2016). Classification with asymmetric label noise: Consistency and maximal denoising. *Electronic Journal of Statistics* **10** 2780–2824.

[2] Blanco, V., Japón, A. and Puerto, J. (2020). A Mathematical Programming approach to Binary Supervised Classification with Label Noise. *arXiv preprint arXiv:2004.10170*.

[3] Bootkrajang, J. and Kabán, A. (2012). Label-noise robust logistic regression and its applications. In *Joint European conference on machine learning and knowledge discovery in databases* 143–158. Springer.

[4] Bootkrajang, J. and Kabán, A. (2014). Learning kernel logistic regression in the presence of class label noise. *Pattern Recognition* **47** 3641–3655.

[5] Cannings, T. I., Fan, Y. and Samworth, R. J. (2020). Classification with imperfect training labels. *Biometrika* **107** 311–330.

[6] Cheng, J., Liu, T., Ramamohanarao, K. and Tao, D. (2020). Learning with Bounded Instance and Label-dependent Label Noise. In *International Conference on Machine Learning* 1789–1799. PMLR.

[7] Frénay, B., Kabán, A. et al. (2014). A comprehensive introduction to label noise. In *ESANN*. Citeseer.

[8] Frenay, B. and Verleysen, M. (2014). Classification in the Presence of Label Noise: A Survey. *IEEE Transactions on Neural Networks and Learning Systems* **25** 845-869.

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

[10] Koltchinskii, V. (2011). *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: Ecole d'Eté de Probabilités de Saint-Flour XXXVIII-2008* **2033**. Springer Science & Business Media.

[11] Kontorovich, A. (2014). Concentration in unbounded metric spaces and algorithmic stability. In *International Conference on Machine Learning* 28–36. PMLR.

[12] Lin, J. Z. and Bradic, J. (2021). Learning to Combat Noisy Labels via Classification Margins. *arXiv preprint arXiv:2102.00751*.

[13] Liu, T. and Tao, D. (2016). Classification with Noisy Labels by Importance Reweighting. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **38** 447–461.
[14] Long, P. M. and Servedio, R. A. (2010). Random classification noise defeats all convex potential boosters. *Machine learning* **78** 287–304.
[15] Lorentz, G. G., Golitschek, M. v. and Makovoz, Y. (1996). *Constructive approximation: advanced problems* **304**. Springer.
[16] Manwani, N. and Sastry, P. (2013). Noise tolerance under risk minimization. *IEEE transactions on cybernetics* **43** 1146–1151.
[17] Menon, A. K., Van Rooyen, B. and Natarajan, N. (2016). Learning from binary labels with instance-dependent corruption. *arXiv preprint arXiv:1605.00751*.
[18] Natarajan, N., Dhillon, I. S., Ravikumar, P. and Tewari, A. (2013). Learning with noisy labels. In *NIPS* **26** 1196–1204.
[19] Natarajan, N., Dhillon, I. S., Ravikumar, P. and Tewari, A. (2018). Cost-Sensitive Learning with Noisy Labels. *Journal of Machine Learning Research* **18** 1-33.
[20] Patrini, G., Nielsen, F., Nock, R. and Carioni, M. (2016). Loss factorization, weakly supervised learning and label noise robustness. In *International conference on machine learning* 708–717. PMLR.
[21] Patrini, G., Rozza, A., Krishna Menon, A., Nock, R. and Qu, L. (2017). Making deep neural networks robust to label noise: A loss correction approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 1944–1952.
[22] Reeve, H. W., Cannings, T. I. and Samworth, R. J. (2021). Adaptive transfer learning. *The Annals of Statistics* **49** 3618–3649.
[23] Reeve, H. et al. (2019). Classification with unknown class-conditional label noise on non-compact feature spaces. In *Conference on Learning Theory* 2624–2651. PMLR.
[24] Reeve, H. and Kabán, A. (2019). Fast rates for a knn classifier robust to unknown asymmetric label noise. In *International Conference on Machine Learning* 5401–5409. PMLR.
[25] Scott, C., Blanchard, G. and Handy, G. (2013). Classification with asymmetric label noise: Consistency and maximal denoising. In *Conference on learning theory* 489–511. PMLR.
[26] Van Rooyen, B., Menon, A. K. and Williamson, R. C. (2015). Learning with symmetric label noise: The importance of being unhinged. *arXiv preprint arXiv:1505.07634*.
[27] van Rooyen, B. and Williamson, R. C. (2018). A Theory of Learning with Corrupted Labels. *Journal of Machine Learning Research* **18** 1-50.