Journal of Machine Learning Research 22 (2021) 1-30 Submitted 9/20; Revised 3/21; Published 5/21

Finite-sample Analysis of Interpolating Linear Classifiers in the Overparameterized Regime

Niladri S. Chatterji CHATTERJI@BERKELEY.EDU
University of California, Berkeley, 366 LeConte Hall, Berkeley, CA 94720

Philip M. Long PLONG@GOOGLE.COM
Google, 1600 Amphitheatre Parkway, Mountain View, CA 94043

Editor: Samory Kpotufe

Abstract

We prove bounds on the population risk of the maximum margin algorithm for two-class linear classification. For linearly separable training data, the maximum margin algorithm has been shown in previous work to be equivalent to a limit of training with logistic loss using gradient descent, as the training error is driven to zero. We analyze this algorithm applied to random data including misclassification noise. Our assumptions on the clean data include the case in which the class-conditional distributions are standard normal distributions. The misclassification noise may be chosen by an adversary, subject to a limit on the fraction of corrupted labels. Our bounds show that, with sufficient overparameterization, the maximum margin algorithm trained on noisy data can achieve nearly optimal population risk.

Keywords: high-dimensional statistics, classification, class-conditional Gaussians, finite-sample analysis, risk bounds

1. Introduction

A surprising statistical phenomenon has emerged in modern machine learning: highly complex models can interpolate training data while still generalizing well to test data, even in the presence of label noise. This is rather striking as it the goes against the grain of the classical statistical wisdom which dictates that predictors that generalize well should trade off between the fit to the training data and the some measure of the complexity or smoothness of the predictor. Many estimators like neural networks, kernel estimators, nearest neighbour estimators, and even linear models have been shown to demonstrate this phenomenon (see, Zhang et al., 2017; Belkin et al., 2019a, among others).

This phenomenon has recently inspired intense theoretical research. One line of work (Soudry et al., 2018; Ji and Telgarsky, 2019; Gunasekar et al., 2017; Nacson et al., 2019; Gunasekar et al., 2018b,a) formalized the argument (Neyshabur et al., 2014; Neyshabur, 2017) that, even when there is no explicit regularization that is used in training these rich models, there is nevertheless implicit regularization encoded in the choice of the optimization method used. For example, in the setting of linear classification, Soudry et al. (2018), Ji and Telgarsky (2019) and Nacson et al. (2019) show that learning a linear classifier using gradient descent on the unregularized logistic or exponential loss asymptotically leads the solution to converge to the maximum ℓ_2-margin classifier. More concretely, given n linearly separable samples $(x_i, y_i)_{j=1}^n$, where $x_i \in \mathbb{R}^p$ are the features and $y_i \in \{-1, 1\}$, the iterates of gradient descent (initialized at the origin) are given by,

$$v^{(t+1)} := v^{(t)} - \alpha \nabla R_{\log}(v^{(t)}) \text{ where } R_{\log}(v) := \sum_{i=1}^{n} \log\left(1 + \exp\left(-y_i(v \cdot x_i)\right)\right).$$

They show that in the large-t limit the normalized predictor obtained by gradient descent $v^{(t)}/\|v^{(t)}\|$ converges to $w/\|w\|$ where,

$$w = \operatorname*{argmin}_{u \in \mathbb{R}^p} \|u\|, \quad (1)$$
$$\text{such that, } y_i(u \cdot x_i) \geq 1, \quad \text{for all } i \in [n].$$

That is, w is the maximum ℓ_2-margin classifier over the training data.

The question still remains, though, why do these maximum margin classifiers generalize well beyond the training set, despite the fact that they "fit the noise"? The fact that $p > n$ renders traditional distribution-free bounds (Cover, 1965; Vapnik, 1982) vacuous. Due to the presence of label noise, margin bounds (Vapnik, 1995; Shawe-Taylor et al., 1998) are also not an obvious answer.

In this paper, we prove an upper bound on the misclassification test error for the maximum margin linear classifier, and therefore on the the limit of gradient descent on the training error without any complexity penalty. Our analysis holds under a natural and fairly general generative model for the data. One special case is where adversarial label noise (see Kearns et al., 1994; Kalai et al., 2008; Klivans et al., 2009; Awasthi et al., 2017; Talwar, 2020) is added to data in which the positive examples are distributed as $\mathsf{N}(\mu, I)$ and the negative examples are distributed $\mathsf{N}(-\mu, I)$. If $\|\mu\|$ is not too small, the clean data will consist of overlapping but largely separate clouds of points. Our assumptions are weaker than this, however (see Section 2 for the details). They are satisfied by the case in which misclassification noise is added to the generative model underlying Fisher's linear discriminant (see Duda et al., 2012; Hastie et al., 2009) (except that, to make the analysis cleaner, the distribution is shifted so that the origin is halfway between the class-conditional means). They also include as special cases the rare-weak model (Donoho and Jin, 2008; Jin, 2009) and a Boolean variant (Helmbold and Long, 2012). We study the overparameterized regime, when the dimension p is significantly greater than the number n of samples. For a precise statement of our main result see Theorem 1. After its statement, we give examples of its consequences, including cases in which s of the p variables are relevant, but weakly associated with the class designations. In some cases where s, p and n are polynomially related, the risk of the maximum-margin algorithm approaches the Bayes-optimal risk as e^{-n^τ}, for $\tau > 0$.

Analysis of classification is hindered by the fact that, in contrast to regression, there is no known simple expression for the parameters as a function of the training data. Our analysis leverages recent results, mentioned above, that characterize the weight vector obtained by minimizing the logistic loss on training data (Soudry et al., 2018; Ji and Telgarsky, 2019; Nacson et al., 2019). We use this result not only to motivate the analysis of the maximum margin algorithm, but also in our proofs, to get a handle on the relationship

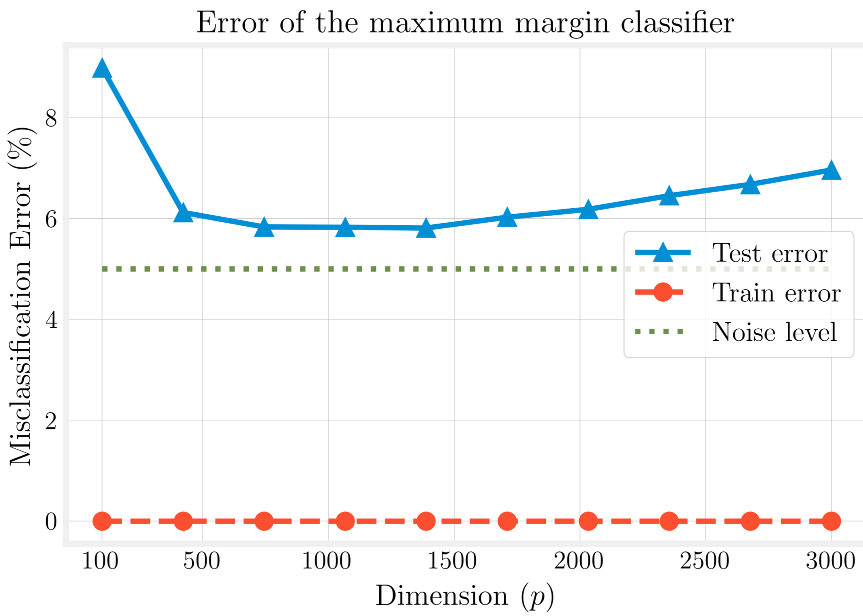

Figure 1: Plot of the test error (solid, blue) and train error (dashed, red) versus the dimension of the covariates p. The number of samples $n = 100$ is kept fixed. The dimension p is varied in the interval $[100, 3000]$. The data is generated according to the Boolean noisy rare-weak model (see Section 2). First, a clean label $\tilde{y}$ is drawn by randomly flipping a fair coin. The covariates x are drawn conditioned on $\tilde{y}$. The first 100 attributes, $(x_1, \ldots, x_{100})$ are equal to the clean label $\tilde{y}$ with probability 0.7, the remaining attributes $(x_{101}, \ldots, x_p)$ are either -1 or 1 with equal probability. The noisy label sample y is generated by flipping the true label $\tilde{y}$ with probability $\eta = 0.05$. The classifier is the maximum ℓ_2-margin classifier defined in Equation (1). The plot is generated by averaging over 500 draws of the samples. The train error on all runs was always 0.

of this solution to the training data. When learning in the presence of label noise, algorithms that minimize a convex loss face the hazard that mislabeled examples can exert an outsized influence. However, we show that in the over-parameterized regime this effect is ameliorated. In particular, we show that the ratio between the (exponential) losses of any two examples is bounded above by an absolute constant. One special case of our upper bounds is where there are relatively few relevant variables, and many irrelevant variables. In this case, classification using only parameters that correctly classify the clean examples with a large margin leads to large loss on examples with noisy class labels. However, the training process can use the parameters on irrelevant variables to play a role akin to slack variables, allowing the algorithm to correctly classify incorrectly labeled training examples with limited harm on independent test data. On the other hand, if there are too many irrelevant variables, accurate classification is impossible (Jin, 2009). Our bounds reflect this reality—if the number of irrelevant variables increases while the number and quality of the relevant variables remains fixed, ultimately our bounds degrade.

In simulation experiments, we see a decrease in population risk with the increase of p beyond n, as observed in previous double-descent papers, but this is followed by an increase. As mentioned above, Jin (2009) showed that, under certain conditions, if the number p of attributes and the number s of relevant attributes satisfy $p \geq s^2$, then, in a sense, learning is impossible. Our experiments suggest that interpolation with logistic loss can succeed close to this boundary, despite the lack of explicit regularization or feature selection.

1.1 Related Work

A number of recent papers have focused on bounding the asymptotic error of overparameterized decision rules. Hastie et al. (2019) and Muthukumar et al. (2020b) studied the asymptotic squared error of the interpolating ordinary least squares estimator for the problem of overparameterized linear regression. This was followed by Mei and Montanari (2019) who characterize the asymptotic error of the OLS estimator in the random features model. As we do, Montanari et al. (2019) studied linear classification, calculating a formula for the asymptotic test error of the maximum margin classifier in the overparameterized setting when the features are generated from a Gaussian distribution and the labels are generated from a logistic link function. This was followed by the work of Liang and Sur (2020) who calculate a formula for the asymptotic test error of the maximum ℓ_1-margin classifier in the same setting. Deng et al. (2020) independently obtained related results, including analysis of the case where the marginal distribution over the covariates is a mixture of Gaussians, one for each class. Previously, Candès and Sur (2020); Sur and Candès (2019) studied the asymptotic test error for this problem in the underparameterized regime (when $p < n$). In contrast with this previous work, we provide finite-sample bounds.

There has also been quite a bit of work on the non-asymptotic analysis of interpolating estimators. Liang and Rakhlin (2020) provided a finite-sample upper bound the expected squared error for kernel "ridgeless" regressor, which interpolates the training data. Kobak et al. (2020) provided an analysis of linear regression that emphasized the role of irrelevant variables as providing placeholders for learning parameters that play the role of slack variables. Belkin et al. (2020) provided a finite-sample analysis of interpolating least-norm regression with feature selection. They showed that, beyond the point where the number of features included in the model exceeds the number of training examples, the excess risk decreases with the number of included features. This analysis considered the case that the covariates have a standard normal distribution. They also obtained similar results for a "random features" model. Bartlett et al. (2020) provided non-asymptotic upper and lower bounds on the squared error for the OLS estimator; their analysis emphasized the effect of the covariance structure of the independent variables on the success or failure of this estimator. This earlier work studied regression; here we consider classification. Study of regression is facilitated by the fact that the OLS parameter vector has a simple closed-form expression as a function of the training data. Belkin et al. (2018) studied the generalization error for a simplicial interpolating nearest neighbor rule. Belkin et al. (2019b) provided bounds on the generalization error for the Nadaraya-Watson regression estimator applied to a singular kernel, a method that interpolates the training data. Liang et al. (2020) provided upper bounds on the population risk for the least-norm interpolant applied to a class of kernels including the Neural Tangent Kernel.

In concurrent independent work, Muthukumar et al. (2020a) studied the generalization properties of the maximum margin classifier in the case where the marginal distribution on the covariates is a single Gaussian, rather than a Gaussian per class. They showed that, in this setting, if there is enough overparameterization, every example is a support vector, so that the maximum margin algorithm outputs the same parameters as the OLS algorithm. They also showed that the accuracy of the model, measured using the 0-1 test loss, can be much better than its accuracy with respect to the quadratic loss.

Additional related work is described in Section 6.

2. Definitions, Notation and Assumptions

Throughout this section, $C > 0$ and $0 < \kappa < 1$ denote absolute constants. We will show that any choice C that is large enough relative to $1/\kappa$ will work.

We study learning from independent random examples $(x, y) \in \mathbb{R}^p \times \{-1, 1\}$ sampled from a joint distribution P. This distribution may be viewed as a noisy variant of another distribution $\tilde{\mathsf{P}}$ which we now describe. A sample from $\tilde{\mathsf{P}}$ may be generated by the following process.

1. First, a clean label $\tilde{y} \in \{-1, 1\}$ is generated by flipping a fair coin.

2. Next, $q \in \mathbb{R}^p$ is sampled from $\mathsf{Q} := \mathsf{Q}_1 \times \cdots \times \mathsf{Q}_p$, which is an arbitrary product distribution over $\mathbb{R}^p$

   - whose marginals are all zero-mean sub-Gaussians with sub-Gaussian norm at most 1 (see Definition 12), and
   - such that $\mathbb{E}_{q \sim \mathsf{Q}}[\|q\|^2] \geq \kappa p$.

3. For an arbitrary unitary matrix U and $\mu \in \mathbb{R}^p$, $x = Uq + \tilde{y}\mu$. This ensures that the mean of x is μ when $\tilde{y} = 1$ and is $-\mu$ when $\tilde{y} = -1$.

4. Finally, noise is modeled as follows. For $0 \leq \eta \leq 1/C$, P is an arbitrary distribution over $\mathbb{R}^p \times \{-1, 1\}$

   - whose marginal distribution on $\mathbb{R}^p$ is the same as $\tilde{\mathsf{P}}$, and
   - such that $d_{TV}(\mathsf{P}, \tilde{\mathsf{P}}) \leq \eta$.

   Note that this definition includes the special case where y is obtained from $\tilde{y}$ by flipping it with probability η.

Choosing a bound of 1 on the sub-Gaussian norm of the components of Q fixes the scale of the data. This simplifies the proofs without materially affecting the analysis, since rescaling the data does not affect the accuracy of the maximum margin algorithm.

Let $(x_1, y_1), \ldots, (x_n, y_n)$ be n training examples drawn according to P. Let

$$\mathcal{S} := \{(x_1, y_1), \ldots, (x_n, y_n)\}.$$

When $\mathcal{S}$ is linearly separable, let $w \in \mathbb{R}^p$ minimize $\|w\|$ subject to $y_1(w \cdot x_1) \geq 1, \ldots, y_n(w \cdot x_n) \geq 1$. (In the setting that we analyze, we will show that, with high probability, $\mathcal{S}$ is linearly separable. When it is not, w may be chosen arbitrarily.)

We will provide bounds on the misclassification probability of the classifier parameterized by w that can be achieved with probability $1-\delta$ over the draw of the samples.

We make the following assumptions on the parameters of the problem:

(A.1) the failure probability satisfies $0 \leq \delta < 1/C$,

(A.2) number of samples satisfies $n \geq C \log(1/\delta)$,

(A.3) the dimension satisfies $p \geq C \max\{\|\mu\|^2 n, n^2 \log(n/\delta)\}$,

(A.4) the norm of the mean satisfies $\|\mu\|^2 \geq C \log(n/\delta)$.

Here are some examples of generative models that fall within our framework.

Example 1 (Gaussian class-conditional model) *The clean labels $\tilde{y}$ are drawn by flipping a fair coin. The distribution on x, after conditioning on the value of $\tilde{y}$, is $\mathsf{N}(\tilde{y}\mu, \Sigma)$, for Σ with $\|\Sigma\| \leq 1$ and $\|\Sigma^{-1}\| \leq 1/\kappa$ (here $\|\Sigma\|$ is the matrix operator norm).*

Example 2 (Noisy rare-weak model) *A special case of the model described above is when $\Sigma = I$ and the mean vector μ is such that only s components are non-zero and all non-zero entries are equal to $\gamma \in \mathbb{R}$.*

Donoho and Jin (2008) studied this model in the noise-free case (i.e. where $\eta = 0$).

Example 3 (Boolean noisy rare-weak model) *Our assumptions are also satisfied*[1] *by the following setting with Boolean attributes.*

- *$\tilde{y} \in \{-1, 1\}$ is generated first, by flipping a fair coin.*
- *For $\gamma \in (0, 1/2)$, the components of $x \in \mathbb{R}^p$ are conditionally independent given $\tilde{y}$: $x_1, \ldots, x_s$ are equal to $\tilde{y}$ with probability $1/2 + \gamma$, $x_{s+1}, \ldots, x_p$ are equal to $\tilde{y}$ with probability $1/2$.*
- *y is obtained from $\tilde{y}$ by flipping it with probability η.*

The noiseless setting of this model was studied by Helmbold and Long (2012).

3. Main Result and Its Consequences

Our main result is a finite-sample bound on the misclassification error of the maximum margin classifier.

Theorem 1 *For all $0 < \kappa < 1$, there is an absolute constant $c > 0$ such that, under the assumptions of Section 2, for all large enough C, with probability $1-\delta$, training on $\mathcal{S}$ produces a maximum margin classifier w satisfying*

$$\mathbb{P}_{(x,y)\sim\mathsf{P}}[\mathrm{sign}(w \cdot x) \neq y] \leq \eta + \exp\left(-c\frac{\|\mu\|^4}{p}\right).$$

1. Strictly speaking, x needs to be scaled down to make the sub-Gaussian norm less than 1 for this to be true, but this does not affect the accuracy of the maximum margin classifier.

Consider the scenario where the number of samples n is a constant, but where the number of dimensions p and $\|\mu\|$ are growing. Then our assumptions require $\|\mu\|^2 = O(p)$.[2] But, for the misclassification error to decrease we need $\|\mu\|^4 = \omega(p)$. Thus if, $\|\mu\| = \Theta(p^\beta)$ for any $\beta \in (1/4, 1/2]$ then as $p \to \infty$, the misclassification error asymptotically will approach the noise level η.

Here are the implications of our results in the noisy rare-weak model. Recall that in this model μ is non-zero only on s coordinates and the non-zero coordinates of μ are equal to some γ. Therefore, $\|\mu\|^2 = \gamma^2 s$.

Corollary 2 *There is an absolute constant $c > 0$ such that, under the assumptions of Section 2, in the noisy rare-weak model, for any $\gamma \geq 0$ and all large enough C, with probability $1 - \delta$, training on $\mathcal{S}$ produces a maximum margin classifier w satisfying*

$$\mathbb{P}_{(x,y)\sim \mathsf{P}}[\text{sign}(w \cdot x) \neq y] \leq \eta + \exp\left(-c\frac{\gamma^4 s^2}{p}\right).$$

Next, let us examine the implications of our results in the Boolean noisy rare-weak model. Here, $\|\mu\|^2 = 4\gamma^2 s$.

Corollary 3 *There is an absolute constant $c > 0$ such that, under the assumptions of Section 2, in the Boolean noisy rare-weak model for any $0 < \gamma < 1/2$ and all large enough C, with probability $1 - \delta$, training on $\mathcal{S}$ produces a maximum margin classifier w satisfying*

$$\mathbb{P}_{(x,y)\sim \mathsf{P}}[\text{sign}(w \cdot x) \neq y] \leq \eta + \exp\left(-c\frac{\gamma^4 s^2}{p}\right).$$

To gain some intuition let us explore the scaling of the misclassification error in these problems in different scaling limits for the parameters in both these problems.

Consider a case where, δ, γ and n are constants and s and p grow. Our assumptions hold if $\|\mu\|^2 = \gamma^2 s = O(p)$. But for the misclassification error to decrease we need $s^2 = \omega(p)$. So if $s = \Theta(p^\beta)$ where, $\beta \in (1/2, 1]$ then the misclassification error scales as $\eta + \exp(-cp^{2\beta-1})$ and asymptotically approaches η.

Jin (2009) showed that for the noiseless rare-weak model learning is impossible when $s = O(\sqrt{p})$ and n is a constant. Our upper bounds show that, in a sense, the maximum margin classifier succeeds arbitrarily close to this threshold.

Another interesting scenario is when δ and γ are constants while both s and p grow as a function of the number of samples n. Let $p = \Theta(n^{2+\rho})$ and $s = \Theta(n^{1+\lambda})$, for positive ρ and λ. Our assumptions are satisfied if $\rho > \lambda$ for large enough n, while, for the misclassification error to reduce with n we need $2\lambda > \rho$. As n gets larger the bound on the misclassification error scales as $\eta + \exp(-cn^{2\lambda-\rho})$ and gets arbitrarily close to η for large enough n. Informally, if the adversary fully expends its noise budget, the Bayes error rate will be at least η; this is true in particular in the case where labels are flipped with probability η. In such cases, even if one could prove that the training data likely to be separated by a large margin, the bound of Theorem 1 approaches the Bayes error rate faster than the standard margin bounds (Vapnik, 1995; Shawe-Taylor et al., 1998).

2. The definitions of "big Oh notation", i.e. $O(\cdot)$, $\omega(\cdot)$, $\Theta(\cdot), \Omega(\cdot)$, may be found in (Cormen et al., 2009).

4. Proof of Theorem 1

First, we may assume without loss of generality that $U = I$. To see this, note that

- if w is the maximum margin classifier for $(x_1, y_1), \ldots, (x_n, y_n)$ then Uw is the maximum margin classifier for $(Ux_1, y_1), \ldots, (Ux_n, y_n)$, and
- the probability that $y(w \cdot x) < 0$ is the same as the probability that $y(Uw \cdot Ux) < 0$.

Let us assume from now on that $U = I$.

Our first lemma is an immediate consequence of the coupling lemma (Lindvall, 2002; Daskalakis, 2011) that allows us to handle the noise in the samples.

Lemma 4 *There is a joint distribution on* $((x, y), (\tilde{x}, \tilde{y}))$ *such that*

- *the marginal on* (x, y) *is* P,
- *the marginal on* $(\tilde{x}, \tilde{y})$ *is* $\tilde{\mathsf{P}}$,
- $\mathbb{P}[x = \tilde{x}] = 1$, *and*
- $\mathbb{P}[y \neq \tilde{y}] \leq \eta$.

Definition 5 *Let* $(x_1, y_1, \tilde{y}_1), \ldots, (x_n, y_n, \tilde{y}_n)$ *be* n *i.i.d. draws from the coupling of Lemma 4, with the redundant* $\tilde{x}_1, \ldots, \tilde{x}_n$ *thrown out. Let* $\mathcal{N}$ *be the set* $\{k : y_k \neq \tilde{y}_k\}$ *of indices of "noisy" examples, and* $\mathcal{C} = \{k : y_k = \tilde{y}_k\}$ *be the indices of "clean" examples.*

Note that Lemma 4 implies that $(x_1, y_1), \ldots, (x_n, y_n)$ are n i.i.d. draws from P, as before.

The next lemma is bound on the misclassification error in terms of the expected value of the *margin* on clean points, $\mathbb{E}_{(x,\tilde{y})\sim\mathsf{P}}[\tilde{y}(w \cdot x)] = \mu \cdot w$, and the norm of the classifier w.

Lemma 6 *There is an absolute positive constant* c *such that*

$$\mathbb{P}_{(x,y)\sim\mathsf{P}}[\text{sign}(w \cdot x) \neq y] \leq \eta + \exp\left(-c\frac{(\mu \cdot w)^2}{\|w\|^2}\right).$$

Proof Observe that

$$\mathbb{P}_{(x,y)\sim\mathsf{P}}[\text{sign}(w \cdot x) \neq y] = \mathbb{P}_{(x,y)\sim\mathsf{P}}[y(w \cdot x) < 0].$$

For a draw $x, y, \tilde{y}$ from the coupling of Lemma 4, we have

$$\begin{aligned}\mathbb{P}[y(w \cdot x) < 0] &\leq \eta + \mathbb{P}[y(w \cdot x) < 0 \text{ and } y = \tilde{y}] \\ &= \eta + \mathbb{P}[\tilde{y}(w \cdot x) < 0].\end{aligned}$$

For $i \leq p$, the ith component of $\tilde{y}x$ is distributed as a $\mu_i + q_i$, where the $q_i \sim \mathsf{Q}_i$ is a mean zero random variable. Thus $\mathbb{E}[\tilde{y}(w \cdot x)] = w \cdot \mu$, so

$$\begin{aligned}\mathbb{P}[\tilde{y}(w \cdot x) < 0] &= \mathbb{P}[\tilde{y}(w \cdot x) - \mathbb{E}[\tilde{y}(w \cdot x)] < -\mu \cdot w] \\ &= \mathbb{P}[w \cdot (\tilde{y}x - \mathbb{E}[\tilde{y}x]) < -\mu \cdot w].\end{aligned}$$

An application of the general Hoeffding's inequality (see Theorem 13) upper bounds this probability and completes the proof. ■

In light of the previous lemma, next we prove a high probability lower bound on the expected margin on a clean point, $\mu \cdot w$.

Lemma 7 *For all $0 < \kappa < 1$, there is an absolute positive constant c such that, for all large enough C, with probability $1-\delta$ over the random choice of $\mathcal{S}$, it is linearly separable, and the maximum margin weight vector w satisfies,*

$$\mu \cdot w \geq \frac{\|w\| \, \|\mu\|^2}{c\sqrt{p}}.$$

Given these two main lemmas above, the main theorem follows immediately.

Proof (of Theorem 1): Combine the result of Lemma 6 with the lower bound on $(\mu \cdot w)$ established in Lemma 7. ■

It remains to prove Lemma 7, a lower bound on the expected margin on clean points $(\mu \cdot w)$. This crucial lemma is proved through a series of auxiliary lemmas, which use a characterization of the maximum margin classifier w in terms of iterates $\{v^{(t)}\}_{t=1}^{\infty}$ of gradient descent on the exponential loss. Denote the risk associated with the exponential loss[3] as

$$R(v) := \sum_{k=1}^{n} \exp\left(-y_k v \cdot x_k\right).$$

Then the iterates of gradient descent are defined as follows:

- $v^{(0)} := 0$, and
- $v^{(t+1)} := v^{(t)} - \alpha \nabla R(v^{(t)})$,

where α is a constant step-size.

Lemma 8 (Soudry et al., 2018, Theorem 3) *For any linearly separable $\mathcal{S}$ and for all small enough step-sizes α, we have*

$$\frac{w}{\|w\|} = \lim_{t \to \infty} \frac{v^{(t)}}{\|v^{(t)}\|}.$$

Definition 9 *For each index k of an example, let $z_k := y_k x_k$.*

Most of the argument required to prove Lemma 6 is deterministic apart from some standard concentration arguments, which are gathered in the following lemma. (Recall that, since we are in the process of proving Theorem 1, the assumptions of Section 2 are in scope.)

Lemma 10 *For all $\kappa > 0$, there is a $c \geq 1$ such that, for all $c' > 0$, for all large enough C, with probability $1-\delta$ over the draw of the samples the following events simultaneously*

3. We could also work with the logistic loss here, but the proofs are simpler if we work with the exponential loss without changing the conclusions.

occur:

$$\text{For all } k \in [n],\ \frac{p}{c} \le \|z_k\|^2 \le cp. \tag{2}$$

$$\text{For all } i \neq j \in [n],\ |z_i \cdot z_j| < c(\|\mu\|^2 + \sqrt{p \log(n/\delta)}). \tag{3}$$

$$\text{For all } k \in \mathcal{C},\ |\mu \cdot z_k - \|\mu\|^2| < \|\mu\|^2/2. \tag{4}$$

$$\text{For all } k \in \mathcal{N},\ |\mu \cdot z_k - (-\|\mu\|^2)| < \|\mu\|^2/2. \tag{5}$$

$$\text{The number of noisy samples satisfies } |\mathcal{N}| \le (\eta + c')n. \tag{6}$$

$$\text{The samples are linearly separable.} \tag{7}$$

The proof of this lemma is in Appendix A.

From here on, we will assume that $\mathcal{S}$ satisfies all the conditions shown to hold with high probability in Lemma 10.

A concern is that, late in training, noisy examples will have outsized effect on the classifier learned. Lemma 11 below limits the extent to which this can be true. It shows that throughout the training process the loss on any one example is at most a constant factor larger than the loss on any other example. This is sufficient since the gradient of the exponential loss

$$\nabla R(v) = -\sum_{k=1}^{n} z_k \exp(-z_k \cdot v),$$

is the sum of the $-z_k$ values weighted by their losses. We also know that with high probability $p/c \le \|z_k\| \le cp$, therefore, showing that the loss on a sample is within a constant factor of the loss of any other sample controls the influence that any one point can have on the learning process. We formalize this intuition in the proof of Lemma 7 in the sequel.

As will be clear in the proof of Lemma 11, the high dimensionality of the classifier (p being larger than $\|\mu\|^2 n$ and $n^2 \log(n/\delta)$) is crucial in showing that the ratio of the losses between any pair of points is bounded. Here is some rough intuition why this is the case.

For the sake of intuition consider the extreme scenario where all the vectors z_k are mutually orthogonal and $\|z_i\| = p$, for all $i \in [n]$. Then in this case, the change in the loss of a sample $i \in [n]$ due to each gradient descent update will be independent of any other sample $j \neq i \in [n]$ and all the losses will decrease exactly at the same rate. Lemma 10 implies that, when p is large enough relative to $\|\mu\|$, the z_k vectors are nearly pairwise orthogonal. In this case, the losses remain within a constant factor of one another.

Lemma 11 *There is an absolute constant c such that, for all large enough C, and all small enough step sizes α, for all iterations $t \ge 0$,*

$$A_t^{max} := \max_{k,\ell \in \mathcal{S}} \left\{ \frac{\exp(-v^{(t)} \cdot z_k)}{\exp(-v^{(t)} \cdot z_\ell)} \right\} \le c.$$

Proof $A_t^{\max}$ is the maximum ratio between a pair of samples at iteration t. Let c_1 be the constant $c \ge 1$ from Lemma 10. We will prove that $A_t^{\max} \le 4c_1^2$ for all $t \ge 0$ by using an inductive argument over the iterations t.

Let us begin by establishing this for the base case, when $t = 0$. Since the gradient descent algorithm is initialized at the origin, the loss for any sample $j \in [n]$ is $\exp(-0 \cdot z_j) = 1$. Therefore, $A_0^{\max} = 1 < 4c_1^2$.

Assume that the inductive hypothesis holds for some iteration t, we shall now prove that then it must also hold at iteration $t + 1$.

To simplify notation we shall analyze the ratio between the losses on the first and the second sample but a similar analysis holds for any distinct pair. Let G_t be the loss on sample z_1 and let H_t be the loss on sample z_2 at the t^{th} iteration. Define $A_t := G_t/H_t$ to be the ratio of the losses at iteration t.

By the definition of $v^{(t+1)}$ as the gradient descent iterate

$$
\begin{aligned}
A_{t+1} &= \frac{\exp\left(-v^{(t+1)} \cdot z_1\right)}{\exp\left(-v^{(t+1)} \cdot z_2\right)} \\
&= \frac{\exp\left(-(v^{(t)} - \alpha \nabla R(v^{(t)})) \cdot z_1\right)}{\exp\left(-(v^{(t)} - \alpha \nabla R(v^{(t)})) \cdot z_2\right)} \\
&= \frac{\exp\left(-v^{(t)} \cdot z_1\right)}{\exp\left(-v^{(t)} \cdot z_2\right)} \cdot \frac{\exp(\alpha \nabla R(v^{(t)}) \cdot z_1)}{\exp(\alpha \nabla R(v^{(t)}) \cdot z_2)} \\
&= A_t \cdot \frac{\exp(-\alpha \sum_{j \in [n]} z_j \cdot z_1 \exp(-v^{(t)} \cdot z_j))}{\exp(-\alpha \sum_{j \in [n]} z_j \cdot z_2 \exp(-v^{(t)} \cdot z_j))} \\
&= A_t \cdot \frac{\exp(-\alpha \|z_1\|^2 G_t)}{\exp(-\alpha \|z_2\|^2 H_t)} \frac{\exp(-\alpha \sum_{j>1} z_j \cdot z_1 \exp(-v^{(t)} \cdot z_j))}{\exp(-\alpha \sum_{j \neq 2} z_j \cdot z_2 \exp(-v^{(t)} \cdot z_j))}.
\end{aligned}
$$

Recalling that c_1 is the constant c from Lemma 10, by (2), we have

$$
\frac{p}{c_1} \leq \|z_i\|^2 \leq c_1 p, \qquad \text{for all } i \in [n],
$$

and (3) gives

$$
|z_i \cdot z_j| < c_1(\|\mu\|^2 + \sqrt{p \log(n/\delta)}), \qquad \text{for all } i \neq j \in [n].
$$

These, combined with the the expression for A_{t+1} above, give

$$
\begin{aligned}
&A_{t+1} \\
&= A_t \cdot \exp(-\alpha \|z_1\|^2 G_t + \alpha \|z_2\|^2 H_t) \cdot \frac{\exp(-\alpha \sum_{j>1} z_j \cdot z_1 \exp(-v^{(t)} \cdot z_j))}{\exp(-\alpha \sum_{j \neq 2} z_j \cdot z_2 \exp(-v^{(t)} \cdot z_j))} \\
&\leq A_t \exp\left(-\alpha p \left(\frac{G_t}{c_1} - c_1 H_t\right)\right) \exp\left(2\alpha c_1 (\|\mu\|^2 + \sqrt{p \log(n/\delta)}) \sum_{j \in [n]} \exp(-v^{(t)} \cdot z_j)\right)
\end{aligned}
$$

which implies

$$
\begin{aligned}
&A_{t+1} \\
&\leq A_t \exp\left(-\frac{\alpha H_t p}{c_1}\left(A_t - c_1^2\right)\right) \exp\left(2\alpha c_1 (\|\mu\|^2 + \sqrt{p \log(n/\delta)}) \sum_{j \in [n]} \exp(-v^{(t)} \cdot z_j)\right). \qquad (8)
\end{aligned}
$$

Consider two disjoint cases.
Case 1 ($A_t \le 2c_1^2$): Using Inequality (8)

$$\begin{aligned}
A_{t+1} &\le A_t \exp\left(-\frac{\alpha H_t p}{c_1}(A_t - c_1^2)\right) \exp\left(2\alpha c_1(\|\mu\|^2 + \sqrt{p\log(n/\delta)}) \sum_{j\in[n]} \exp(-v^{(t)}\cdot z_j)\right) \\
&\le A_t \exp\left(c_1 \alpha H_t p\right) \exp\left(2\alpha c_1(\|\mu\|^2 + \sqrt{p\log(n/\delta)}) \sum_{j\in[n]} \exp(-v^{(t)}\cdot z_j)\right) \\
&\overset{(i)}{\le} A_t \exp\left(c_1 \alpha p n\right) \exp\left(2\alpha c_1(\|\mu\|^2 + \sqrt{p\log(n/\delta)})n\right) \\
&= A_t \exp\left(\alpha(c_1 p + 2c_1(\|\mu\|^2 + \sqrt{p\log(n/\delta)}))n\right) \\
&\overset{(ii)}{\le} 2c_1^2 \exp(1/8) < 4c_1^2
\end{aligned}$$

where (i) follows since the sum of the losses on all samples is always smaller than the initial loss which is n (see Lemma 22) and $H_t \le n$, while, (ii) follows as the step-size may be chosen to be at most $(8c_1(p + 2(\|\mu\|^2 + \sqrt{p\log(n/\delta)})\|\mu\|^2)n)^{-1}$.
Case 2 ($A_t > 2c_1^2$) : Reusing Inequality (8),

$$\begin{aligned}
A_{t+1} &\le A_t \exp\left(-\frac{\alpha H_t p}{c_1}\left(A_t - c_1^2\right)\right) \exp\left(2\alpha c_1(\|\mu\|^2 + \sqrt{p\log(n/\delta)}) \sum_{j\in[n]} \exp(-v^{(t)}\cdot z_j)\right) \\
&= A_t \exp\left(-\frac{\alpha H_t p}{c_1}\left(A_t - c_1^2\right)\right) \exp\left(2\alpha c_1(\|\mu\|^2 + \sqrt{p\log(n/\delta)}) H_t \sum_{j\in[n]} \frac{\exp(-v^{(t)}\cdot z_j)}{H_t}\right) \\
&\le A_t \exp\left(-\frac{\alpha H_t p}{c_1}\left(A_t - c_1^2\right)\right) \exp\left(8\alpha c_1^3(\|\mu\|^2 + \sqrt{p\log(n/\delta)}) H_t n\right) \quad \text{(by the IH)} \\
&= A_t \exp\left(-\alpha H_t \left(\frac{p}{c_1}\left(A_t - c_1^2\right) - 8c_1^3(\|\mu\|^2 + \sqrt{p\log(n/\delta)})n\right)\right) \\
&\le A_t \exp\left(-\alpha H_t \left(c_1 p - 8c_1^3(\|\mu\|^2 + \sqrt{p\log(n/\delta)})n\right)\right).
\end{aligned}$$

Since $p > C\|\mu\|^2$ and $p > Cn^2\log(n/\delta)$, and noting that Lemma 10 is consistent with C being arbitrarily large while c_1 remains fixed, we have that, in this case, $A_{t+1} \le A_t$ (as the term in the exponent is non-positive). This completes the proof of the inductive step in this case, and therefore the entire proof.

■

4.1 Proof of Lemma 7

Armed with Lemma 11, we now prove Lemma 7.

Let us proceed assuming that the event defined in Lemma 10 occurs, and, in this proof, let c_1 be the constant c from that lemma. We know that this event occurs with probability at least $1 - \delta$.

We have

$$\mu \cdot v^{(t+1)} = \mu \cdot v^{(t)} + \alpha \sum_{k=1}^{n} (\mu \cdot z_k) \exp\left(-v^{(t)} \cdot z_k\right).$$

Dividing the sum into the clean and noisy examples, we have

$$\begin{aligned} \mu \cdot v^{(t+1)} = \mu \cdot v^{(t)} + \alpha \sum_{k \in \mathcal{C}} (\mu \cdot z_k) \exp\left(-v^{(t)} \cdot z_k\right) \\ + \alpha \sum_{k \in \mathcal{N}} (\mu \cdot z_k) \exp\left(-v^{(t)} \cdot z_k\right). \end{aligned} \tag{9}$$

Combining (4), (5) and (9) we infer

$$\begin{aligned} \mu \cdot v^{(t+1)} &\geq \mu \cdot v^{(t)} + \frac{\|\mu\|^2 \alpha}{2} \sum_{k \in \mathcal{C}} \exp\left(-v^{(t)} \cdot z_k\right) - \frac{3\|\mu\|^2 \alpha}{2} \sum_{k \in \mathcal{N}} \exp\left(-v^{(t)} \cdot z_k\right) \\ &= \mu \cdot v^{(t)} + \frac{\|\mu\|^2}{2} \alpha R(v^{(t)}) - 2\|\mu\|^2 \alpha \sum_{k \in \mathcal{N}} \exp\left(-v^{(t)} \cdot z_k\right). \end{aligned} \tag{10}$$

Since $|\mathcal{N}| \leq (\eta + c')n$, where c' is an arbitrarily small constant, if c_2 is the constant from Lemma 11, we have

$$\sum_{k \in \mathcal{N}} \exp\left(-v^{(t)} \cdot z_k\right) \leq c_2(\eta + c')n \min_k \exp\left(-v^{(t)} \cdot z_k\right) \leq c_2(\eta + c')R(v^{(t)}) \leq R(v^{(t)})/4,$$

since $\eta \leq 1/C$. Thus Inequality (10) implies

$$\mu \cdot v^{(t+1)} \geq \mu \cdot v^{(t)} + \frac{\|\mu\|^2 \alpha}{4} R(v^{(t)}).$$

Unrolling this via an induction yields

$$\mu \cdot v^{(t+1)} \geq \frac{\alpha \|\mu\|^2}{4} \sum_{m=0}^{t} R(v^{(m)}) \qquad (\text{since } v^{(0)} = 0).$$

Now let us multiply both sides by $\|w\|/\|v_{t+1}\|$

$$\|w\| \frac{\mu \cdot v^{(t+1)}}{\|v^{(t+1)}\|} \geq \|w\| \frac{\alpha \|\mu\|^2 \sum_{m=0}^{t} R(v^{(m)})}{4\|v^{(t+1)}\|}.$$

Next, let us take the large-t limit. Applying Lemma 8 to the left hand side,

$$\mu \cdot w \geq \alpha \|w\| \|\mu\|^2 \lim_{t \to \infty} \frac{\sum_{m=0}^{t} R(v^{(m)})}{4\|v^{(t+1)}\|}. \tag{11}$$

By definition of the gradient descent iterates

$$\begin{aligned}\|v^{(t+1)}\| &= \left\|\sum_{m=0}^{t} \alpha \nabla R(v^{(m)})\right\| \\ &\le \alpha \sum_{m=0}^{t} \|\nabla R(v^{(m)})\| \\ &= \alpha \sum_{m=0}^{t} \left\|\sum_{k=1}^{n} z_k \exp(-v^{(m)} \cdot z_k)\right\| \\ &\le \alpha \sum_{m=0}^{t} \sum_{k=1}^{n} \exp(-v^{(m)} \cdot z_k)\|z_k\| \\ &\le \alpha c_1 \sqrt{p} \sum_{m=0}^{t} R(v^{(m)}).\end{aligned}$$

This together with Inequality (11) yields

$$\mu \cdot w \ge \frac{\|w\|\|\mu\|^2}{4c_1\sqrt{p}},$$

completing the proof.

5. Simulations

We experimentally study the behavior of the maximum margin classifier in the overparameterized regime on synthetic data generated according to the Boolean noisy rare-weak model. Recall that this is a model where the clean label $\tilde{y} \in \{-1, 1\}$ is first generated by flipping a fair coin. Then the covariate x is drawn from a distribution conditioned on $\tilde{y}$ such that s of the coordinates of x are equal to $\tilde{y}$ with probability $1/2 + \gamma$ and the other $p - s$ coordinates are random and independent of the true label. The noisy label y is obtained by flipping $\tilde{y}$ with probability η. In this section the flipping probability η is always 0.1. In all our experiments the number of samples n is kept constant at 100.

In the first experiment in Figure 2 we hold n and the number of relevant attributes s constant and vary the dimension p for different values of γ. We find that after an initial dip in the test error (for $\gamma = 0.2, 0.3$) the test error starts to rise slowly with p, as in our upper bounds.

Next, in Figure 3 we explore the scaling of the test error with the number of relevant attributes s when n and p are held constant. As we would expect, the test error decreases as s grows for all the different values of γ.

Finally, in Figure 4 we study how the test error changes when both p and s are increasing when n and γ are held constant. Our results (see Corollary 3) do not guarantee learning when $s = \Theta(\sqrt{p})$ (and Jin (2009) proved that learning is impossible in a related setting, even in the absence of noise); we find that the test error remains constant in our experiment in this setting. In the cases when $s = p^{0.55}$ and when $s = p^{0.65}$, slightly beyond this threshold, the test error approaches the Bayes-optimal error as p gets large in our experiment.

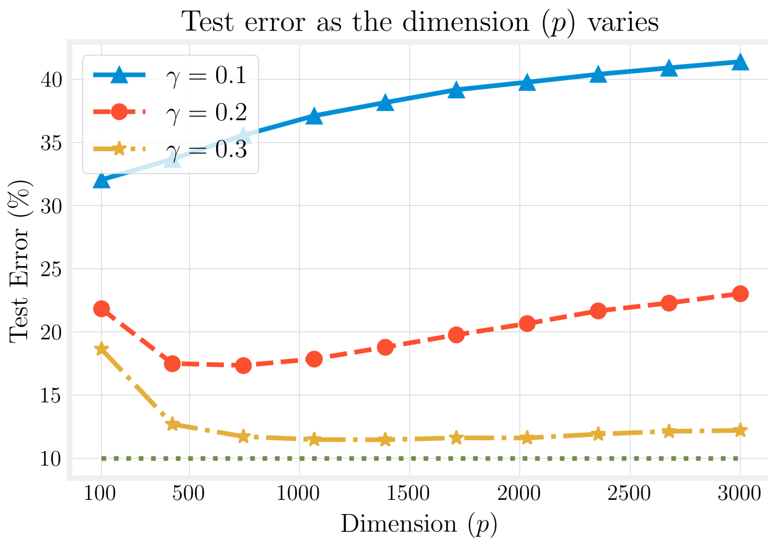

Figure 2: Plot of the test error versus the dimension of the covariates p for different values of γ. The number of samples $n = 100$ and the number of relevant variables $s = 50$ are both kept fixed. The dimension p is varied in the interval $[100, 3000]$. The data is generated according to the Boolean noisy rare-weak model. The dotted olive green line represents the noise level (10%). The plot is generated by averaging over 500 draws of the samples. The train error on all runs was always 0.

This provides experimental evidence that the maximum margin algorithm, without explicit regularization or feature selection, even in the presence of noise, learns with using a number of relevant variables near the theoretical limit of what is possible for any algorithm. (Note that, as emphasized by Helmbold and Long, 2012, the fraction of relevant variables is going to zero as p increases in these experiments.)

6. Additional Related Work

Ng and Jordan (2002) compared the Naive Bayes algorithm, which builds a classifier from estimates of class-conditional distributions using conditional independence assumptions, with discriminative training of a logistic regressor. Their main point is that Naive Bayes converges faster. Our analysis provides a counterpoint to theirs, showing that, for a reasonable data distribution that includes label noise, in the overparameterized regime, unregularized discriminative training with a commonly used loss function learns a highly accurate classifier from a constant number of examples.

Analysis of learning with class-conditional Gaussians with the same covariance structure has been extensively studied in the case that the number n of training examples is greater than the number p of parameters (see Anderson, 2003). When $p \gg n$, Bickel and Levina (2004) showed that, even when the class-conditional distributions do not have diagonal covariance matrices, behaving as if they are can lead to improved classification accuracy.

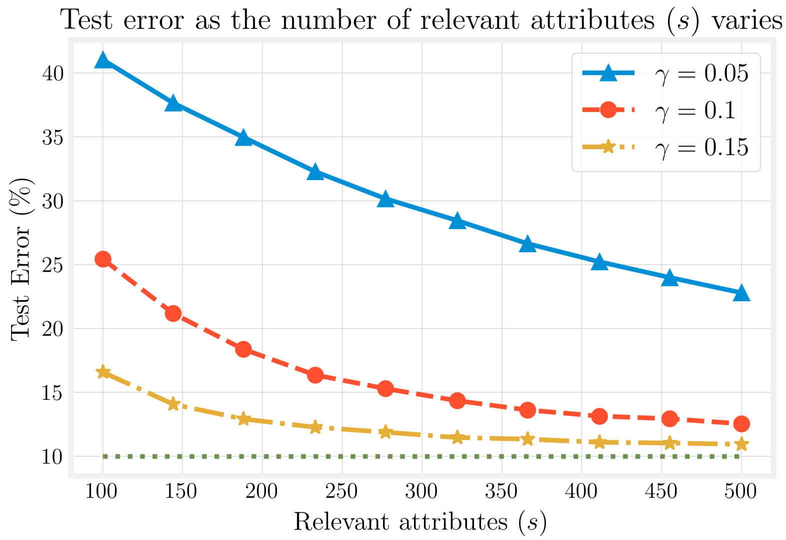

Figure 3: Plot of the test error versus the number of relevant attributes s for different values of γ. The number of samples $n = 100$ and the dimension $p = 500$ are both kept fixed. The dimension s is varied in the interval $[100, 500]$. The data is generated according to the Boolean noisy rare-weak model. The dotted olive green line represents the noise level (10%). The plot is generated by averaging over 500 draws of the samples. The train error on all runs was always 0.

The model that we use is a generalization of the rare-weak model studied by Donoho and Jin (2008) (see Section 2 for details of our set-up). The class-conditional distributions studied there have a standard multivariate normal distribution, while our results hold for a more general class of sub-Gaussian class-conditional distributions. More importantly, in order to address the experimental findings of Zhang et al. (2017), we have have also supplemented the rare-weak model to include label noise. Finite-sample bounds for algorithms using ℓ_1 penalties, again, in the absence of label noise have been obtained (Cai and Liu, 2011; Li et al., 2015; Li and Jia, 2017; Cai and Zhang, 2019). Dobriban and Wager (2018) studied regularized classification in the asymptotic framework where p and n go to infinity together. Fan and Fan (2008) and Jin (2009) proved that learning with class-conditional Gaussians is impossible when too few variables are associated with the class designations. Our analysis shows that, even in the presence of misclassification noise, in a sense, the maximum-margin algorithm succeeds up to the edge of the frontier established by one of the results by Jin (2009). Nagarajan and Kolter (2019) used a data distribution like the distributions considered here, but to analyze limitations of uniform convergence tools.

The framework studied here also includes as a special case the setting studied by Helmbold and Long (2012), with Boolean attributes; again, a key modification is the addition of misclassification noise. Also, while the upper bounds of Helmbold and Long (2012) are for algorithms that perform unweighted votes over selected attributes, here we consider the maximum margin algorithm. A more refined analysis of learning with conditionally indepen-

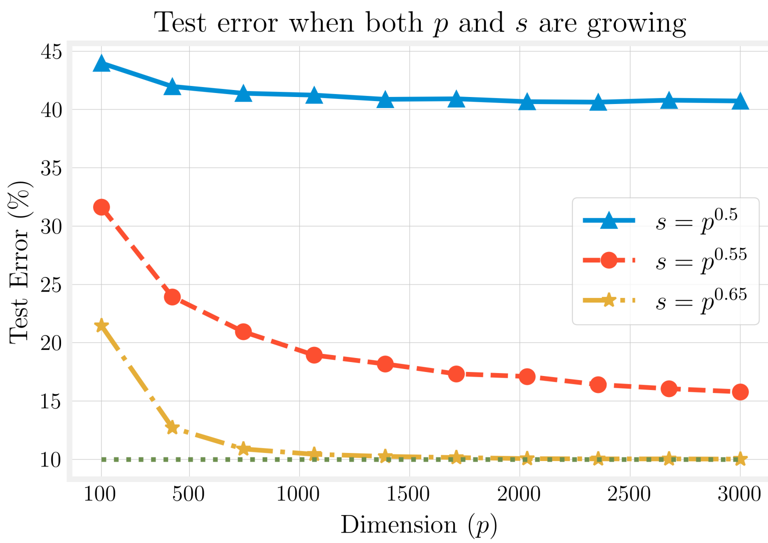

Figure 4: Plot of the test error versus the dimension (p) for different scalings of s with p. The number of samples $n = 100$ and $\gamma = 0.1$ are both held fixed. The dimension p is varied in the interval $[100, 3000]$. The data is generated according to the Boolean noisy rare-weak model. The dotted olive green line represents the noise level (10%). The plot is generated by averaging over 500 draws of the samples. The train error on all runs was always 0.

dent Boolean attributes was carried out by Berend and Kontorovich (2015). Kleindessner and Awasthi (2018) studied learning with conditionally independent Boolean attributes in the presence of noise—they analyzed tasks other than classification, including estimating the degree of association between the attributes (viewed in that work as experts) and the true class designations.

As mentioned above, we consider the case that the data is corrupted with label noise. We consider adversarial label noise (Kearns et al., 1994; Kalai et al., 2008; Klivans et al., 2009; Awasthi et al., 2017; Talwar, 2020). In this model, an adversary is allowed to change the classifications of an arbitrary subset of the domain whose probability is η, while leaving the marginal distribution on the covariates unchanged. It includes as a special case the heavily studied situation in which classifications are randomly flipped with probability η (Angluin and Laird, 1988; Kearns, 1998; Cesa-Bianchi et al., 1999; Servedio, 1999; Kalai and Servedio, 2005; Long and Servedio, 2010; Van Rooyen et al., 2015) along with variants that allow limited dependence of the probability that a label is corrupted on the clean example (Lugosi, 1992; Natarajan et al., 2013; Scott et al., 2013; Cannings et al., 2020). Adversarial label noise allows for the possibility that noise is concentrated in a part of the domain, where noisy examples have greater potential to coordinate their effects; it is a weaker assumption than even Massart noise (Massart and Nédélec, 2006; Blanchard et al., 2008; Awasthi et al., 2015; Diakonikolas et al., 2019), which requires a separate limit on the conditional probability of an incorrect label, given any clean example. We show that, with sufficient overparameterization, even in the absence of regularity in the noise, the algorithm that simply minimizes the standard softmax loss without any explicit regularization enjoys surprisingly strong noise tolerance.

After a preliminary version of this paper was posted on arXiv (Chatterji and Long, 2020), some related work was published (Wang and Thrampoulidis, 2020; Hsu et al., 2021; Liang and Recht, 2021).

7. Discussion

Even in the presence of misclassification noise, with sufficient overparameterization, unregularized minimization of the logistic loss produces accurate classifiers when the clean data has well-separated sub-Gaussian class-conditional distributions.

We have analyzed the case of a linear classifier without a bias term. In the setting studied here, the Bayes-optimal classifier has a bias term of zero, and adding analysis of a bias term in the maximum margin classifier would complicate the analysis without significantly changing the results.

In the noisy rare-weak model, when p and s scale favorably with n, and γ is a constant, the excess risk of the maximum margin algorithm decreases very rapidly with n. One contributing cause is a "wisdom of the crowds" effect that is present when classifying with conditionally independent attributes: a classifier can be very accurate, even when the angle between its normal vector and the optimum is not very small. For example, if 100 experts each predict a binary class, and they are correct independently with probability $3/4$, a vote over their predictions remains over 95% accurate even if we flip the votes of 25 of them. (Note that, even in some cases where Lemma 7 implies accuracy very close to optimal, it may not imply that the cosine of the angle between μ and w is anywhere near 1.) On the other hand, the concentration required for successful generalization is robust to departures from the conditional independence assumption. Our assumptions already allow substantial class-conditional dependence among the attributes, but it may be interesting to explore even weaker assumptions.

We note that a bound on the accuracy of the maximum margin classifier with respect to the distribution $\tilde{\mathsf{P}}$ without any label noise is implicit in our analysis. (The bound is the same as Theorem 1, but without the η.)

Our bounds show that the maximum margin classifier approaches the Bayes risk as the parameters go to infinity in various ways. It would be interesting to characterize the conditions under which this happens. A related question is to prove lower bounds in terms of the parameters of the problem. Another is to prove bounds for finite p and n under weaker conditions.

We assume that the distributions of $x - \mu$ and $x - (-\mu)$ are the same. This is useful in particular for simplifying the analysis of dot products between examples of opposite classes. It should not be difficult to extend the analysis meaningfully to remove this assumption—we use this assumption in this paper to keep the analysis as simple and clean as possible.

The distribution of $x - \tilde{y}\mu$ comes from a sub-Gaussian distribution obtained by applying a unitary transformation to a latent variable with a product distribution. Concentration theorems have been proved under many conditions weaker than independence (see Schmidt et al., 1995; Dubhashi and Ranjan, 1998; Pemmaraju, 2001). Our analysis can be straightforwardly extended using these weaker assumptions.

The assumption that the latent variable q satisfies $\mathbb{E}_{q\sim Q}[\|q\|^2] \geq \kappa p$ formalizes the notion that, in a sense, these variables are truly used, which is required for concentration. We believe that smaller values of $\mathbb{E}_{q\sim Q}[\|q\|^2]$ are compatible with successful learning. We chose this assumption to facilitate a simple and clean analysis, but an analysis that separates the dependence on $\mathbb{E}_{q\sim Q}[\|q\|^2]$ is a potential subject for future work.

The lower bounds on p are needed for concentration, as described earlier. We suspect that the requirement that $p = \Omega(n^2 \log(n/\delta))$ can be improved. The bottleneck is in the proof of Lemma 11. (As we mentioned earlier, a larger value of p promotes the property that the loss on an example can be reduced by gradient descent without increasing the loss on other examples very much.)

The lower bound on $\|\mu\|^2$ allows us to focus on the case where most clean examples are classified correctly by a large margin, which is the case that we want to focus on. This could potentially be weakened or removed through a case analysis, exploiting the fact that a weaker bound is needed in the case that $\|\mu\|$ is small.

Using the generalized Hoeffding bound (Theorem 13) it is not hard to show (see Helmbold and Long, 2012, Theorem 1) that, in our setting, the Bayes optimal classifier has error at most $\eta + \exp(-c\|\mu\|^2)$ for an absolute constant c, and Slud's Lemma gives a similar lower bound (see Anthony and Bartlett, 2009) (see also Helmbold and Long, 2012, Inequality (8)). Our upper bound of $\eta + \exp(-c'\|\mu\|^4/p)$ for the maximum margin algorithm applied to finite training data is worse than this by a factor of $\|\mu\|^2/p$ in the exponent.

Implicit regularization lemmas like the one that was so helpful to us have been obtained for other problems (see Gunasekar et al., 2017, 2018b; Woodworth et al., 2020; Azulay et al., 2021). We hope that further advances in implicit regularization research could be combined with the techniques of this paper to prove generalization guarantees for interpolating classifiers using richer model classes, including neural networks.

Acknowledgments

We thank anonymous reviewers for their valuable comments and suggestions. NC gratefully acknowledges the support of the NSF through grants IIS-1619362 and IIS-1909365.

Appendix A. Concentration Inequalities

In this section we begin by presenting a definition of sub-Gaussian and sub-exponential random variables in terms of Orlicz norms. Then we state a version of Hoeffding's inequality and a version of Bernstein's inequality. Finally, we prove Lemma 10 which implies that a good event which our proofs rely on holds with high probability.

For an excellent reference of sub-Gaussian and sub-exponential concentration inequalities we refer the reader to Vershynin (2018, Chapter 2).

Definition 12 (sub-Gaussian random variable) *A random variable θ is sub-Gaussian if*

$$\|\theta\|_{\psi_2} := \inf\left\{t > 0 : \mathbb{E}[\exp(\theta^2/t^2)] < 2\right\}$$

is bounded. Further, $\|\theta\|_{\psi_2}$ is defined to be its sub-Gaussian norm.

We now state general Hoeffding's inequality (see Vershynin, 2018, Theorem 2.6.3) a concentration inequality for a weighted sum of independent sub-Gaussian random variables.

Theorem 13 (General Hoeffding's inequality) *Let $\theta_1, \ldots, \theta_m$ be independent mean-zero sub-Gaussian random variables and $a = (a_1, \ldots, a_m) \in \mathbb{R}^m$. Then, for every $t > 0$, we have*

$$\mathbb{P}\left[\Big|\sum_{i=1}^{m} a_i\theta_i\Big| \geq t\right] \leq 2\exp\left(-\frac{c_2 t^2}{K^2\|a\|^2}\right),$$

where $K = \max_i \|\theta_i\|_{\psi_2}$ and c_2 is an absolute constant.

A one-sided version of this theorem (upper/lower deviation bound) holds without the factor of 2 multiplying the exponent on the right hand side.

Definition 14 (sub-exponential random variable) *A random variable θ is said to be sub-exponential if*

$$\|\theta\|_{\psi_1} := \inf\left\{t > 0 : \mathbb{E}[\exp(|\theta|/t) < 2]\right\}$$

is bounded. Further, $\|\theta\|_{\psi_1}$ is defined to be its sub-exponential norm.

We shall also use Bernstein's inequality (see Vershynin, 2018, Theorem 2.8.1) a concentration inequality for a sum of independent sub-exponential random variables.

Theorem 15 (Bernstein's inequality) *For independent mean-zero sub-exponential random variables $\theta_1, \ldots, \theta_m$, for every $t > 0$, we have*

$$\mathbb{P}\left[\Big|\sum_{i=1}^{m} \theta_i\Big| \geq t\right] \leq 2\exp\left(-c_1 \min\left\{\frac{t^2}{\sum_{i=1}^{m}\|\theta_i\|_{\psi_1}^2}, \frac{t}{\max_i\|\theta_i\|_{\psi_1}}\right\}\right),$$

where c_1 is an absolute constant.

Again note that a one-sided version of this inequality holds without the factor of 2 multiplying the exponent on the right hand side.

We break the proof of Lemma 10 into different parts, which are proved in separate lemmas. Lemma 10 then follows by a union bound.

Lemma 16 *For all $\kappa > 0$, there is a $c \geq 1$ such that, for all large enough C, with probability at least $1 - \delta/6$, for all $k \in [n]$,*

$$\frac{p}{c} \leq \|z_k\|^2 \leq cp.$$

Proof For any *clean* sample z_i, the random variables $(z_{ij} - \mu_j)^2$ are sub-exponential with norm

$$\|(z_{ij} - \mu_j)^2\|_{\psi_1} \leq \|z_{ij} - \mu_j\|_{\psi_2}^2 \leq 1.$$

After centering, the sub-exponential norm of the zero-mean random variable $(z_{ij} - \mu_j)^2 - \mathbb{E}[(z_{ij} - \mu_j)^2]$ is at most a constant (see Vershynin, 2018, Exercise 2.7.10). Therefore, by Bernstein's inequality, there is an absolute constant c_0 such that

$$\mathbb{P}[|\|z_i - \mu\|_2^2 - \mathbb{E}[\|z_i - \mu\|_2^2]| \geq t] \leq 2\exp\left(-c_0 \min\left\{\frac{t^2}{p}, t\right\}\right).$$

By setting $t = \kappa p/2$

$$\mathbb{P}[|\|z_i - \mu\|_2^2 - \mathbb{E}[\|z_i - \mu\|_2^2]| \geq \kappa p/2] \leq \frac{\delta}{6n},$$

since $p \geq C\log(n/\delta)$ for a large constant C. Recall that by assumption, $\kappa p \leq \mathbb{E}[\|z_i - \mu\|_2^2] \leq 3p$, where the upper bound follows from the assumption that the components of q have sub-Gaussian norm at most 1. Recalling that $\kappa \leq 1$,

$$\frac{\kappa p}{2} \leq \|z_i - \mu\|^2 \leq 4p \tag{12}$$

with probability at least $1 - \delta/(6n)$.

By Young's inequality for products, $\|z_i - \mu\|^2 \leq 2\|z_i\|^2 + 2\|\mu\|^2$. Also recall that by assumption $\|\mu\|^2 < p/C$. Combining this with the left hand side in the display above, for large enough C, we have

$$\|z_i\|^2 \geq \frac{1}{2}\left(\frac{\kappa p}{2} - 2\|\mu\|^2\right) > \frac{\kappa p}{8}.$$

Again by Young's inequality $\|z_i\|^2 = \|z_i - \mu + \mu\|^2 \leq 2\|z_i - \mu\|^2 + 2\|\mu\|^2$. Therefore,

$$\|z_i\|^2 \leq 2\|z_i - \mu\|^2 + 2\|\mu\|^2 \leq 8p + 2\|\mu\|^2 < 10p,$$

with the same probability.

A similar argument also holds for all noisy samples by considering the random variables $(z_k - (-\mu))$. Hence, by taking a union bound over all samples completes the proof. ■

Lemma 17 *There is a $c \geq 1$ such that, for all large enough C, with probability at least $1 - \delta/6$, for all $i \neq j \in [n]$,*

$$|z_i \cdot z_j| < c(\|\mu\|^2 + \sqrt{p\log(n/\delta)}).$$

Proof First, let us condition on the division of $\{1, \ldots, n\}$ into clean points $\mathcal{C}$ and noisy points $\mathcal{N}$. After this, for each $i \in \mathcal{C}$, $\mathbb{E}[z_i] = \mu$ (where the expectation is conditioned on z_i

being clean), and for each $i \in \mathcal{N}$, $\mathbb{E}[z_i] = -\mu$. For each i, let $\xi_i := z_i - \mathbb{E}[z_i]$. The same logic that proved (12), together with a union bound, yields

$$\mathbb{P}\left[\exists i \in [n],\ \|\xi_j\| > 2\sqrt{p}\right] \leq \frac{\delta}{24}. \tag{13}$$

For any pair $i, j \in [n]$ of indices of examples, we have

$$\mathbb{P}\left[|\xi_i \cdot \xi_j| \geq t\right] \leq \mathbb{P}\big[|\xi_i \cdot \xi_j| \geq t \;\big|\; \|\xi_j\| \leq 2\sqrt{p}\big] + \mathbb{P}\left[\|\xi_j\| > 2\sqrt{p}\right]. \tag{14}$$

If we regard ξ_j as fixed, and only ξ_i as random, Theorem 13 gives

$$\mathbb{P}\big[|\xi_i \cdot \xi_j| \geq t\big] \leq 2\exp\left(-c_2 \frac{t^2}{\|\xi_j\|^2}\right).$$

Thus,

$$\mathbb{P}\big[|\xi_i \cdot \xi_j| \geq t \;\big|\; \|\xi_j\| \leq 2\sqrt{p}\big] \leq 2\exp\left(-c_2 \frac{t^2}{4p}\right) = 2\exp\left(-c_3 \frac{t^2}{p}\right)$$

for $c_3 = c_2/4$. Substituting into Inequality (14), we infer

$$\mathbb{P}\left[|\xi_i \cdot \xi_j| \geq t\right] \leq 2\exp\left(-c_3 \frac{t^2}{p}\right) + \mathbb{P}\left[\|\xi_j\| > 2\sqrt{p}\right].$$

Taking a union bound over all pairs for the first term, and all individuals for the second term, we get

$$\mathbb{P}\left[\exists i \neq j \in [n],\ |\xi_i \cdot \xi_j| \geq t\right] \leq 2n^2 \exp\left(-c_3 \frac{t^2}{p}\right) + \mathbb{P}\left[\exists j \in [n],\ \|\xi_j\| > 2\sqrt{p}\right].$$

Choosing $t = c\sqrt{p\log(n/\delta)}$ for a large enough value of c, we have

$$\mathbb{P}\left[\exists i \neq j \in [n],\ |\xi_i \cdot \xi_j| \geq c\sqrt{p\log(n/\delta)}\right] \leq \frac{\delta}{24} + \mathbb{P}\left[\exists j \in [n],\ \|\xi_j\| > 2\sqrt{p}\right].$$

Applying (13), we get

$$\mathbb{P}\left[\exists i \neq j \in [n],\ |\xi_i \cdot \xi_j| \geq c\sqrt{p\log(n/\delta)}\right] \leq \frac{\delta}{12}. \tag{15}$$

For any i, clean or noisy, Hoeffding's inequality implies

$$\mathbb{P}\left[|\mu \cdot z_i| > 2\|\mu\|^2\right] < 2\exp\left(-c_2 \frac{4\|\mu\|^4}{\|\mu\|^2}\right) = 2\exp\left(-4c_2 c^2 \|\mu\|^2\right).$$

Since $\|\mu\|^2 \geq C\log(n/\delta)$, this implies

$$\mathbb{P}\left[|\mu \cdot z_i| > 2\|\mu\|^2\right] < \frac{\delta}{12n}.$$

Therefore, by taking a union bound over all $i \in \{1, \ldots, n\}$

$$\mathbb{P}\left[\exists i,\ |\mu \cdot z_i| > 2\|\mu\|^2\right] < \delta/12. \tag{16}$$

Both the events in (15) and (16) will simultaneously hold with probability at most $\delta/6$. Assume that the event complementary to this bad event occurs, then for any distinct pair z_i and z_j

$$\begin{aligned}|z_i \cdot z_j| &= \left|(z_i - \mathbb{E}[z_i]) \cdot (z_j - \mathbb{E}[z_j]) + \mathbb{E}[z_i] \cdot \mathbb{E}[z_j] + \mu \cdot z_i + \mu \cdot z_j\right| \\ &= |\xi_i \cdot \xi_j + \mathbb{E}[z_i] \cdot \mathbb{E}[z_j] + \mu \cdot z_i + \mu \cdot z_j| \\ &\le |\xi_i \cdot \xi_j| + \|\mu\|^2 + |\mu \cdot z_i| + |\mu \cdot z_j| \\ &\le 5\|\mu\|^2 + c\sqrt{p \log(n/\delta)},\end{aligned}$$

which completes our proof. ■

Lemma 18 *For all large enough C, with probability at least $1 - \delta/6$,*

$$\text{for all } k \in \mathcal{C},\ |\mu \cdot z_k - \|\mu\|^2| < \|\mu\|^2/2.$$

Proof If z_k is a clean point then, $\mathbb{E}[z_k | k \in \mathcal{C}] = \mu$. Therefore the random variable $|\mu \cdot z_k - \|\mu\|^2| = |\mu \cdot (z_k - \mu)|$ has sub-Gaussian norm at most $\|\mu\|$. By applying Hoeffding's inequality,

$$\mathbb{P}[|\mu \cdot z_k - \|\mu\|^2| \ge \|\mu\|^2/2] \le \frac{\delta}{6n},$$

since $\|\mu\|^2 > C \log(n/\delta)$. Taking a union bound over all clean points establishes the claim. ■

Lemma 19 *For all large enough C, with probability at least $1 - \delta/6$,*

$$\text{for all } k \in \mathcal{N},\ |\mu \cdot z_k - (-\|\mu\|^2)| < \|\mu\|^2/2.$$

Proof The proof is the same as the proof of Lemma 18, except that for any noisy sample, z_k, the conditional mean $\mathbb{E}[z_k | k \in \mathcal{N}] = -\mu$. ■

Lemma 20 *For all $c' > 0$, for all large enough C, with probability $1 - \delta/6$ the number of noisy samples satisfies $|\mathcal{N}| \le (\eta + c')n$.*

Proof Since $n \ge C \log(1/\delta)$, this follows from a Hoeffding bound. ■

Lemma 21 *If (2) and (3) hold, then, if C is large enough, $(x_1, y_1), \ldots, (x_n, y_n)$ are linearly separable.*

Proof Let $v := \sum_{k=1}^{n} z_k$. For each k and any $\delta > 0$,

$$\begin{aligned} y_k v \cdot x_k &= \sum_i y_i y_k x_i \cdot x_k \\ &\geq p/c_1 - \sum_{i \neq k} y_i y_k x_i \cdot x_k \\ &\geq p/c_1 - c_1 n(\|\mu\|^2 + \sqrt{p \log(n/\delta)}) \\ &> 0 \end{aligned}$$

for $p \geq C \max\{\|\mu\|^2 n, n^2 \log(n/\delta)\}$, completing the proof. ■

Appendix B. Decreasing Loss

Lemma 22 *For all small enough step sizes α, for all iterations t, $R(v^{(t)}) \leq n$.*

Proof Since $R(v^{(0)}) = \sum_{j \in [n]} \exp(0 \cdot z_j) = n$, it suffices to prove that, for all t, $R(v^{(t+1)}) \leq R(v^{(t)})$. Toward showing this, note that, if c_1 is the constant c from Lemma 10, the operator norm of the Hessian at any solution v may be bound as follows:

$$\begin{aligned} \|\nabla^2 R(v)\| &= \left\| \sum_k z_k z_k^\top \exp(-v z_k) \right\| \\ &\leq \sum_k \left\| z_k z_k^\top \right\| \exp(-v z_k) \\ &\leq c_1 p \sum_k \exp(-v z_k) \\ &= c_1 p R(v). \end{aligned}$$

This implies that R is $c_1 pn$-smooth over those v such that $R(v) \leq n$. This implies that, for $\alpha \leq (c_1 pn)^{-1}$, if $R(v^{(t)}) \leq n$ then $R(v^{(t+1)}) \leq R(v^{(t)}) \leq n$ (this can be seen, for example, by mirroring the argument used in Lemma B.2 in Ji and Telgarsky, 2019). The lemma then follows using induction. ■

References

T.W. Anderson. *An introduction to multivariate statistical analysis.* Wiley Series in Probability and Statistics. Wiley, 2003.

Dana Angluin and Philip Laird. Learning from noisy examples. *Machine Learning*, 2(4): 343–370, 1988.

Martin Anthony and Peter L Bartlett. *Neural network learning: Theoretical foundations.* Cambridge University Press, 2009.

Pranjal Awasthi, Maria-Florina Balcan, Nika Haghtalab, and Ruth Urner. Efficient learning of linear separators under bounded noise. In *Conference on Learning Theory*, pages 167–190, 2015.

Pranjal Awasthi, Maria Florina Balcan, and Philip M Long. The power of localization for efficiently learning linear separators with noise. *Journal of the ACM (JACM)*, 63(6):1–27, 2017.

Shahar Azulay, Edward Moroshko, Mor Shpigel Nacson, Blake Woodworth, Nathan Srebro, Amir Globerson, and Daniel Soudry. On the implicit bias of initialization shape: Beyond infinitesimal mirror descent. *arXiv preprint arXiv:2102.09769*, 2021.

Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

Mikhail Belkin, Daniel J Hsu, and Partha Mitra. Overfitting or perfect fitting? Risk bounds for classification and regression rules that interpolate. In *Advances in Neural Information Processing Systems*, pages 2300–2311, 2018.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019a.

Mikhail Belkin, Alexander Rakhlin, and Alexandre B Tsybakov. Does data interpolation contradict statistical optimality? In *International Conference on Artificial Intelligence and Statistics*, pages 1611–1619, 2019b.

Mikhail Belkin, Daniel Hsu, and Ji Xu. Two models of double descent for weak features. *SIAM Journal on Mathematics of Data Science*, 2(4):1167–1180, 2020.

Daniel Berend and Aryeh Kontorovich. A finite sample analysis of the Naive Bayes classifier. *Journal of Machine Learning Research*, 16(44):1519–1545, 2015.

Peter J Bickel and Elizaveta Levina. Some theory for Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations. *Bernoulli*, 10(6):989–1010, 2004.

Gilles Blanchard, Olivier Bousquet, and Pascal Massart. Statistical performance of support vector machines. *The Annals of Statistics*, 36(2):489–531, 2008.

Tony Cai and Weidong Liu. A direct estimation approach to sparse linear discriminant analysis. *Journal of the American statistical association*, 106(496):1566–1577, 2011.

Tony Cai and Linjun Zhang. High dimensional linear discriminant analysis: optimality, adaptive algorithm and missing data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(4):675–705, 2019.

Emmanuel J Candès and Pragya Sur. The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression. *The Annals of Statistics*, 48 (1):27–42, 2020.

Timothy I Cannings, Yingying Fan, and Richard J Samworth. Classification with imperfect training labels. *Biometrika*, 107(2):311–330, 2020.

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

Niladri S Chatterji and Philip M Long. Finite-sample analysis of interpolating linear classifiers in the overparameterized regime. *arXiv preprint arXiv:2004.12019*, 2020.

Thomas H Cormen, Charles E Leiserson, Ronald L Rivest, and Clifford Stein. *Introduction to algorithms*. MIT Press, 2009.

Thomas M Cover. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE transactions on electronic computers*, (3): 326–334, 1965.

Constantinos Daskalakis. Probability and computation, lecture 3. `http://people.csail.mit.edu/costis/6896sp11/lec3s.pdf`, 2011. Scribes: A. Chiesa and Z. Zhu. (Downloaded 4/13/20).

Zeyu Deng, Abla Kammoun, and Christos Thrampoulidis. A model of double descent for high-dimensional logistic regression. In *ICASSP*, pages 4267–4271, 2020.

Ilias Diakonikolas, Themis Gouleakis, and Christos Tzamos. Distribution-independent PAC learning of halfspaces with Massart noise. In *Advances in Neural Information Processing Systems*, pages 4749–4760, 2019.

Edgar Dobriban and Stefan Wager. High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics*, 46(1):247–279, 2018.

David Donoho and Jiashun Jin. Higher criticism thresholding: Optimal feature selection when useful features are rare and weak. *Proceedings of the National Academy of Sciences*, 105(39):14790–14795, 2008.

Devdatt Dubhashi and Desh Ranjan. Balls and bins: A study in negative dependence. *Random Structures & Algorithms*, 13(5):99–124, 1998.

Richard O Duda, Peter E Hart, and David G Stork. *Pattern classification*. John Wiley & Sons, 2012.

Jianqing Fan and Yingying Fan. High dimensional classification using features annealed independence rules. *The Annals of Statistics*, 36(6):2605, 2008.

Suriya Gunasekar, Blake E Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nathna Srebro. Implicit regularization in matrix factorization. In *Advances in Neural Information Processing Systems*, pages 6151–6159, 2017.

Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In *International Conference on Machine Learning*, pages 1832–1841, 2018a.

Suriya Gunasekar, Jason D Lee, Daniel Soudry, and Nathan Srebro. Implicit bias of gradient descent on linear convolutional networks. In *Advances in Neural Information Processing Systems*, pages 9461–9471, 2018b.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*, 2019.

David P Helmbold and Philip M Long. On the necessity of irrelevant variables. *Journal of Machine Learning Research*, 13(Jul):2145–2170, 2012.

Daniel Hsu, Vidya Muthukumar, and Ji Xu. On the proliferation of support vectors in high dimensions. In *International Conference on Artificial Intelligence and Statistics*, pages 91–99, 2021.

Ziwei Ji and Matus Telgarsky. The implicit bias of gradient descent on nonseparable data. In *Conference on Learning Theory*, pages 1772–1798, 2019.

Jiashun Jin. Impossibility of successful classification when useful features are rare and weak. *Proceedings of the National Academy of Sciences*, 106(22):8859–8864, 2009.

Adam Tauman Kalai and Rocco A Servedio. Boosting in the presence of noise. *Journal of Computer and System Sciences*, 71(3):266–290, 2005.

Adam Tauman Kalai, Adam R Klivans, Yishay Mansour, and Rocco A Servedio. Agnostically learning halfspaces. *SIAM Journal on Computing*, 37(6):1777–1805, 2008.

Michael Kearns. Efficient noise-tolerant learning from statistical queries. *Journal of the ACM (JACM)*, 45(6):983–1006, 1998.

Michael J Kearns, Robert E Schapire, and Linda M Sellie. Toward efficient agnostic learning. *Machine Learning*, 17(2-3):115–141, 1994.

Matthäus Kleindessner and Pranjal Awasthi. Crowdsourcing with arbitrary adversaries. In *International Conference on Machine Learning*, pages 2708–2717, 2018.

Adam R Klivans, Philip M Long, and Rocco A Servedio. Learning halfspaces with malicious noise. *Journal of Machine Learning Research*, 10(Dec):2715–2740, 2009.

Dmitry Kobak, Jonathan Lomond, and Benoit Sanchez. The optimal ridge penalty for real-world high-dimensional data can be zero or negative due to the implicit ridge regularization. *Journal of Machine Learning Research*, 21(169):1–16, 2020.

Tianyang Li, Adarsh Prasad, and Pradeep K Ravikumar. Fast classification rates for high-dimensional Gaussian generative models. In *Advances in Neural Information Processing Systems*, pages 1054–1062, 2015.

Yanfang Li and Jinzhu Jia. L1 least squares for sparse high-dimensional LDA. *Electronic Journal of Statistics*, 11(1):2499–2518, 2017.

Tengyuan Liang and Alexander Rakhlin. Just interpolate: Kernel "ridgeless" regression can generalize. *The Annals of Statistics*, 48(3):1329–1347, 2020.

Tengyuan Liang and Benjamin Recht. Interpolating classifiers make few mistakes. *arXiv preprint arXiv:2101.11815*, 2021.

Tengyuan Liang and Pragya Sur. A precise high-dimensional asymptotic theory for boosting and min-ℓ_1-norm interpolated classifiers. *arXiv preprint arXiv:2002.01586*, 2020.

Tengyuan Liang, Alexander Rakhlin, and Xiyu Zhai. On the multiple descent of minimum-norm interpolants and restricted lower isometry of kernels. In *Conference on Learning Theory*, pages 2683–2711, 2020.

Torgny Lindvall. *Lectures on the coupling method.* Courier Corporation, 2002.

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

Gabor Lugosi. Learning with an unreliable teacher. *Pattern Recognition*, 25(1):79–87, 1992.

Pascal Massart and Élodie Nédélec. Risk bounds for statistical learning. *The Annals of Statistics*, 34(5):2326–2366, 2006.

Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and double descent curve. *arXiv preprint arXiv:1908.05355*, 2019.

Andrea Montanari, Feng Ruan, Youngtak Sohn, and Jun Yan. The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime. *arXiv preprint arXiv:1911.01544*, 2019.

Vidya Muthukumar, Adhyyan Narang, Vignesh Subramanian, Mikhail Belkin, Daniel Hsu, and Anant Sahai. Classification vs regression in overparameterized regimes: Does the loss function matter? *arXiv preprint arXiv:2005.08054*, 2020a.

Vidya Muthukumar, Kailas Vodrahalli, Vignesh Subramanian, and Anant Sahai. Harmless interpolation of noisy data in regression. *IEEE Journal on Selected Areas in Information Theory*, 2020b.

Mor Shpigel Nacson, Nathan Srebro, and Daniel Soudry. Stochastic gradient descent on separable data: Exact convergence with a fixed learning rate. In *International Conference on Artificial Intelligence and Statistics*, pages 3051–3059, 2019.

Vaishnavh Nagarajan and J Zico Kolter. Uniform convergence may be unable to explain generalization in deep learning. In *Advances in Neural Information Processing Systems*, pages 11615–11626, 2019.

Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. In *Advances in neural information processing systems*, pages 1196–1204, 2013.

Behnam Neyshabur. Implicit regularization in deep learning. *arXiv preprint arXiv:1709.01953*, 2017.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. *arXiv preprint arXiv:1412.6614*, 2014.

Andrew Y Ng and Michael I Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes. In *Advances in Neural Information Processing Systems*, pages 841–848, 2002.

Sriram V Pemmaraju. Equitable coloring extends Chernoff-Hoeffding bounds. In *Approximation, Randomization, and Combinatorial Optimization: Algorithms and Techniques*, pages 285–296. Springer, 2001.

Jeanette P Schmidt, Alan Siegel, and Aravind Srinivasan. Chernoff–Hoeffding bounds for applications with limited independence. *SIAM Journal on Discrete Mathematics*, 8(2): 223–250, 1995.

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

Rocco A Servedio. On PAC learning using Winnow, Perceptron, and a Perceptron-like algorithm. In *Conference on Computational Learning Theory*, pages 296–307, 1999.

John Shawe-Taylor, Peter L Bartlett, Robert C Williamson, and Martin Anthony. Structural risk minimization over data-dependent hierarchies. *IEEE transactions on Information Theory*, 44(5):1926–1940, 1998.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(1):2822–2878, 2018.

Pragya Sur and Emmanuel J Candès. A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, 116 (29):14516–14525, 2019.

Kunal Talwar. On the error resistance of hinge-loss minimization. In *Advances in Neural Information Processing Systems*, volume 33, pages 4223–4234, 2020.

Brendan Van Rooyen, Aditya Menon, and Robert C Williamson. Learning with symmetric label noise: The importance of being unhinged. In *Advances in Neural Information Processing Systems*, pages 10–18, 2015.

Vladimir Vapnik. *Estimation of dependences based on empirical data.* Springer-Verlag, 1982.

Vladimir N Vapnik. *The nature of statistical learning theory.* Springer-Verlag, 1995.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018.

Ke Wang and Christos Thrampoulidis. Benign overfitting in binary classification of Gaussian mixtures. *arXiv preprint arXiv:2011.09148*, 2020.

Blake Woodworth, Suriya Gunasekar, Jason D Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models. In *Conference on Learning Theory*, pages 3635–3673, 2020.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017.