*The Annals of Statistics*
2022, Vol. 50, No. 3, 1669–1695
https://doi.org/10.1214/22-AOS2170


# A PRECISE HIGH-DIMENSIONAL ASYMPTOTIC THEORY FOR BOOSTING AND MINIMUM-$\ell_1$-NORM INTERPOLATED CLASSIFIERS

BY TENGYUAN LIANG[1,a] AND PRAGYA SUR[*2,b]

[1]*Booth School of Business, University of Chicago,* [a]*tengyuan.liang@chicagobooth.edu*
[2]*Department of Statistics, Harvard University,* [b]*pragya@fas.harvard.edu*

This paper establishes a precise high-dimensional asymptotic theory for boosting on separable data, taking statistical and computational perspectives. We consider a high-dimensional setting where the number of features (weak learners) $p$ scales with the sample size $n$, in an overparametrized regime. Under a class of statistical models, we provide an exact analysis of the generalization error of boosting when the algorithm interpolates the training data and maximizes the empirical $\ell_1$-margin. Further, we explicitly pin down the relation between the boosting test error and the optimal Bayes error, as well as the proportion of active features at interpolation (with zero initialization). In turn, these precise characterizations answer certain questions raised in (*Neural Comput.* **11** (1999) 1493–1517; *Ann. Statist.* **26** (1998) 1651–1686) surrounding boosting, under assumed data generating processes. At the heart of our theory lies an in-depth study of the maximum-$\ell_1$-margin, which can be accurately described by a new system of nonlinear equations; to analyze this margin, we rely on Gaussian comparison techniques and develop a novel uniform deviation argument. Our statistical and computational arguments can handle (1) any finite-rank spiked covariance model for the feature distribution and (2) variants of boosting corresponding to general $\ell_q$-geometry, $q \in [1, 2]$. As a final component, via the Lindeberg principle, we establish a universality result showcasing that the scaled $\ell_1$-margin (asymptotically) remains the same, whether the covariates used for boosting arise from a nonlinear random feature model or an appropriately linearized model with matching moments.

**1. Introduction.** Modern machine learning methods are regularly used for classification tasks. Typically, these algorithms are complex, and often produce solutions with zero training error, even for random labels. Prominent examples include ensemble learning, neural networks and kernel machines. However, among the many solutions that interpolate the training data, not all exhibit superior generalization. Empirically, it has been commonly observed that practical algorithms—running even on large overparametrized models—favor minimal ways of interpolating the training data, which has been conjectured to be crucial for good generalization. Different problem formulations and optimization algorithms favor distinct notions of minimalism, typically measured by specific norms of the classifier. This paper focuses on the celebrated boosting/AdaBoost algorithm in this minimum-norm interpolation regime, where we conduct a precise analysis of its statistical and computational properties under specific data-generating mechanisms.

Ensemble learning algorithms, recognized as powerful toolkits at the disposal of a data scientist, have found widespread usage across domains. Boosting is arguably one of the most powerful ensemble learning algorithms that combines weak learners using intelligent schemes and exhibits remarkable generalization performance. The groundbreaking AdaBoost

Received July 2021; revised December 2021.
[*]Author names are in alphabetical order.
*MSC2020 subject classifications.* Primary 68Q32; secondary 62H30.
*Key words and phrases.* Boosting, high-dimensional asymptotics, minimum-norm interpolation, overparametrization, margin theory.

paper, Freund and Schapire [41], is widely regarded as the milestone in the boosting literature, which can be traced back even earlier [40, 87]. AdaBoost is an iterative algorithm that updates the weights on the training examples adaptively based on the errors incurred at prior iterations. AdaBoost demonstrated preferable generalization capabilities over existing algorithms such as bagging [88], which led to decades of research activities devoted to a better understanding of this algorithm and its variants.

The seminal papers [15, 34, 77] observed that AdaBoost achieves zero error on the training data within a few iterations, whereas the generalization error continues to decrease well beyond this interpolation timepoint. Recently, similar phenomena and puzzles resurfaced in the context of neural networks [106], and motivated the study of interpolation and implicit regularization [5, 8, 10, 51, 63, 64]. This peculiar and seemingly counterintuitive phenomenon naturally piqued the interest of a broad community of statisticians and machine learners. Several explanations emerged over the past two decades.

*Margin-based analyses*. In a breakthrough work, Schapire, Freund, Bartlett and Lee [88] proposed that the generalization performance of the algorithm is crucially tied to a measure of confidence in classification, that can be captured through the (normalized) empirical margin of the training examples. [88] observed that over the course of iterations, AdaBoost creates classifiers such that the fraction of training examples with a large margin increases, and the empirical margin distribution stabilizes to a limiting one rapidly. In particular, given any margin level $\kappa > 0$, they discovered upper bounds on the prediction error that reveal interesting tradeoffs between two terms: (i) the fraction of training examples with margin below $\kappa$, and (ii) the term $\kappa^{-1}\mathrm{C}(\mathcal{H})/\sqrt{n}$ that involves the complexity of the class $\mathrm{C}(\mathcal{H})$ and the sample size $n$ scaled by $\kappa$. A large empirical margin distribution was then conjectured to be a key factor behind the superior generalization performance of certain classifiers. These upper bounds provided extremely useful insights, nonetheless, [88] commented that the proposed upper bounds can be suboptimal in general, and that *"an important open problem is to derive more careful and precise bounds ... Besides paying closer attention to constant factors, such an analysis might also involve the measurement of more sophisticated statistics."* Breiman [16] subsequently contended these empirical margin distribution based explanations, using extensive simulations, and proposed to bound the generalization error using the *minimum value of the margin* over the training set. Later, Koltchinskii and Panchenko [59] improved the earlier bounds from [88]. Despite significant progress in this direction, since these results involved upper bounds, the qualitative question regarding key quantities that precisely determine the generalization behavior of AdaBoost remained unanswered.

*Consistency and early stopping*. In conjunction with the generalization error, statisticians and learning theorists deeply care about the consistency of AdaBoost, and in particular, about the precise relationship between the test error and the optimal Bayes error. The problem of consistency was posed by Breiman [17], who studied convergence properties of the algorithm in the population case. The seminal papers Jiang [56], Lugosi and Vayatis [69], Zhang [107], Koltchinskii and Besnozova [58] considered different function classes and variants of boosting, and furthered this direction of research. [56] established that AdaBoost is process consistent, in the sense that, there exists a stopping time at which the prediction error approximates the optimal Bayes error in the limit of large samples. A parallel understanding emerged from empirical studies conducted in [44, 49, 73, 79]—AdaBoost may overfit, particularly in complex model classes and high noise settings, when left to run for an arbitrary large number of steps. On the one hand, these naturally inspired subsequent work on appropriate regularization strategies for "early stopping" as in Zhang and Yu [108], Bartlett and Traskin [6]. On the other hand, as the model classes become complex and overparametrized, the test error of

Boosting Algorithms may deviate from the optimal Bayes error. Despite an extensive bulk of work, a precise characterization of the test error and its relation to the Bayes error for the overparametrized case is still missing in the current literature.

*Connections with min-$\ell_1$-norm interpolation (and implications).* In a venture to understand the path of boosting iterates better, Rosset, Zhu and Hastie [82], Zhang and Yu [108] established that for linearly separable data, AdaBoost with infinitesimal step size converges to the minimum-$\ell_1$-norm interpolated classifier (equation (1.2)) when left to run forever. This interpolant is crucially related to the maximum $\ell_1$-margin on the data, $\kappa_{n,\ell_1}$ (equation (1.3)). In fact, expressed differently, these results establish that the number of optimization steps necessary for AdaBoost to reach zero training error can be upper bounded by $O(\kappa_{n,\ell_1}^{-2})$. Together with the earlier results Breiman [16], this leads to a plausible conjecture that the max-$\ell_1$-margin is a crucial quantity that determines both generalization and optimization behaviors of Boosting Algorithms. (See also [98], for methods to shrink step sizes so that AdaBoost produces approximate maximum margin classifiers.) Thus, understanding the precise value of this margin, and the iteration time necessary for convergence to the min-$\ell_1$-norm interpolant (on separable data) is crucial for settling such a conjecture. Furthermore, refined analyses of such quantities for various overparametrized models is expected to shed light on the effects of overparametrization on optimization, an understanding that has so far eluded the literature.

Rosset et al. [82] further discussed that the aforementioned convergence to min-$\ell_1$-norm interpolated classifiers indicates the following: boosting potentially converges (in direction) to a sparse classifier. It would then be of interest to understand properties of the limiting solution better, for example, the analyst may wish to understand the number of weak learners deemed important by the boosting solution. This is particularly crucial in today's context where producing interpretable classifiers in high-stakes decision making has critical social consequences [27, 57, 68, 83, 105]. Boosting has subsequently witnessed widespread development, and varying perspectives have emerged through several seminal works, for example, [18, 21, 39, 44, 45, 84]; see Section 4 for further discussions.

*This paper.* Prior literature suggested that the min-$\ell_1$-norm interpolated classifier and the max-$\ell_1$-margin may form central characters behind Boosting Algorithms on linearly separable data. However, a thorough understanding of their exact relations with the boosting solution, whether these are key quantities, and how these objects behave, have so far been lacking. When there is label noise in $y$, conditional on the features $x$, linear separability only happens in an overparametrized regime where the number of features $p$ grows with the sample size $n$; to see this, note that a fixed $p$-dimensional linear model class, cannot shatter $n$-points with all possible signs when $n$ grows.

Furthermore, boosting has empirically demonstrated exceptional performance with many weak-learners. Therefore, to study properties of boosting on separable data, it is both theoretically necessary and empirically natural to analyze the algorithm in a high-dimensional (overparametrized) setting. This paper studies these crucial questions surrounding AdaBoost, in high dimensions, focusing on the case of binary classifications. Our theoretical contributions apply under specific data generating schemes detailed in Sections 2 and 3.5. Throughout the paper, boosting/*Boosting Algorithms* loosely refers to the version of AdaBoost described in Section 2.

To describe our contributions, imagine that we observe $n$ i.i.d. samples $(x_i, y_i)$ drawn from some joint distribution, with $x_i \in \mathbb{R}^p$ abstracting the vector of weak-learners, and labels $y_i \in \{+1, -1\}$. We seek to characterize various properties of boosting in a high-dimensional

setting, and to capture a regime where $p$ is comparable to $n$, assume that $p$ diverges with $n$ at some fixed ratio

$$p/n \to \psi > 0. \tag{1.1}$$

This is a natural high-dimensional setting for analyzing separable data [24, 75], as argued above; this regime has also been investigated for regression problems and other contexts (see, for instance, [31, 32, 35–37, 95, 96, 104], and the references cited therein) and is well known to produce asymptotic predictions with accurate finite sample performance. Since we are primarily interested in overparametrized settings, we assume that the data is (asymptotically) linearly separable in the sense of equation (2.6). This is equivalent to the dimensionality $\psi$ lying above a threshold that depends on the underlying signal strength of the problem [24, 29, 75]; see Section 2 for further details. Define the *min-$\ell_1$-norm interpolated classifier* to be

$$\hat{\theta}_{n,\ell_1} \in \arg\min_{\theta} \|\theta\|_1, \quad \text{s.t. } y_i x_i^\top \theta \geq 1, 1 \leq i \leq n. \tag{1.2}$$

Note that at a finite sample level the min-$\ell_1$-norm interpolants may not be unique, and our asymptotic theory works for any such $\hat{\theta}_{n,\ell_1}$. It is not hard to see that the $\hat{\theta}_{n,\ell_1}$ direction solves the following *max-$\ell_1$-margin* problem

$$\kappa_{n,\ell_1} := \max_{\|\theta\|_1 \leq 1} \min_{1 \leq i \leq n} y_i x_i^\top \theta, \tag{1.3}$$

whenever $\kappa_{n,\ell_1}$ is positive. We first study a stylized model where each row of the design matrix follows a Gaussian distribution with a diagonal covariance, the response is binary and the distribution of the response conditional on the covariates is given by a generalized linear model as in (2.1) (see Section 2 for further details). Later, Section 3.5 presents extensions to showcase that the precise asymptotic theory carries over to spiked covariance models and random feature models. In the aforementioned setting, this paper provides the following contributions to the statistical and computational understanding of boosting:

(i) We provide a precise characterization of the value of the max-$\ell_1$-margin (Theorem 3.1) in the high-dimensional regime (1.1). Informally, we show that $\sqrt{p}\kappa_{n,\ell_1}$ converges almost surely to a constant $\kappa_\star$ that depends on $\psi$ and other problem parameters, such as the signal-to-noise ratio in the data generating model. Theorem 3.1 explicitly pins down the limiting constant $\kappa_\star$; in fact, this can be entirely described by the fixed points of a complicated yet easy to solve nonlinear system of equations that we will introduce in (3.9). This limiting characterization will prove crucial for understanding the properties of boosting on (asymptotically) separable data.

(ii) We establish a precise formula for the generalization error of the min-$\ell_1$-norm interpolant $\hat{\theta}_{n,\ell_1}$ (Theorem 3.2), once again in the regime (1.1). The formula illuminates that the generalization error is completely governed by the dimensionality parameter $\psi$ and the limit $\kappa_\star$ characterized in the preceding step. The consequences of this result for boosting will be discussed soon; notably, the min-$\ell_1$-norm interpolant has been conjectured to be crucial in other contexts (see Section 4) and, therefore, we expect Theorem 3.2 to be of wider importance beyond boosting.

(iii) Turning to boosting, we develop an exact characterization of a threshold $T$ such that for all iterations $t \geq T$, the boosting iterates (with a properly scaled step size) stay arbitrarily close to $\hat{\theta}_{n,\ell_1}$, in the large $n$, $p$ limit (1.1) (Theorem 3.3). This characterization builds upon existing works on margin maximization that provide a $1/\sqrt{t}$ rate [38, 98], and uses the well-known rescaling technique, shrinkage technique and mirror descent connections of boosting (see [26, 38, 50, 54, 89, 108] for a nonexhaustive set of related works). Together with Theorems 3.1–3.2, this result provides an exact characterization of the generalization error of

boosting, and improves upon the existing upper bounds [59, 88], in our setting. Crucially, this formula involves $\kappa_\star$ (through an implicit nonlinear function) and, therefore, our results imply that, at least under the aforementioned data-generation scheme, the max-$\ell_1$-margin drives the generalization performance of boosting. Furthermore, the formula encodes a concrete recipe for comparing the test error of boosting with the Bayes error in high dimensions.

As an aside, we remark that the aforementioned $1/\sqrt{t}$ margin maximization rate has been improved for gradient descent and $\ell_2$ margin maximization [54]. The argument here relies on the smoothness of the dual objective function (also an $\ell_2$ norm), which is absent in our case. This suggests an interesting difference between the $\ell_1$ and $\ell_2$ cases.

(iv) The iteration threshold $T$ from the prior step can be described through a formula (in the large $n$, $p$ limit) that involves the limit of the max-$\ell_1$-margin $\kappa_\star$. Utilizing this, we demonstrate two curious phenomena regarding overparametrization: (1) Keeping other problem parameters fixed, $T$ decreases with an increase in $\psi$, suggesting that *overparametrization helps in optimization*. (2) We establish bounds on the fraction of activated coordinates in the boosting solution (with zero initialization) when it first interpolates the training data.

(v) Finally, we introduce a new class of Boosting Algorithms that converge to the max-$\ell_q$-margin direction (Section 3.4) for $q > 1$. [82] discussed the importance of studying such notions of margins, since it is unclear which geometry induces a better solution (see also [50]). Here, we construct such algorithms and provide precise analyses of their generalization (for the case $q \in [1, 2]$) and optimization properties (for all $q > 1$) in a spirit similar to that for boosting done above.

On the theoretical end, our analyses for the above contributions build upon classical results in Gaussian comparison inequalities [47, 48] that have been strengthened relatively recently [91, 100, 101], leading to the *Convex Gaussian Min-Max Theorem* (CGMT) (see Section 4 for a discussion). The topic of max-$\ell_2$-margin has received considerable attention, dating back to [46, 90], and has more recently been analyzed in [29, 75]. Our proofs begin from these existing theory surrounding the max-$\ell_2$-margin, particularly [29, 75], however, the $\ell_2$ (coordinate invariant) and $\ell_q$ ($q \neq 2$, coordinate specific) geometries differ significantly. Therefore, considerable theoretical work is necessary to obtain the precise characterizations outlined above; our key contributions in this regard are highlighted in Appendix A (see Supplementary Material [66]). Specifically, we introduce a novel uniform deviation argument, which later (Section 3.5) allows us to extend our results to settings with nondiagonal covariance between features. Further elaboration on the difference between the $\ell_1$ and $\ell_2$ cases can be found in Appendix B.

The aforementioned contributions rely on a specific data-generating scheme that might appear stylized. However, the qualitative message remains the same in several settings beyond this specific scheme. Section 3.5 explores this in further detail. In particular, we establish similar characterization for the max-$\ell_1$-margin and the min-$\ell_1$-norm interpolant for a class of models with feature covariances given by spiked covariance matrices (see Section 3.5.1 and Appendix D.1). Our result can be utilized to establish boosting properties analogous to point (iii) above, for these other data generation schemes. We remark that the simplest model in this class—the rank-one perturbation model—corresponds to the standard Gaussian mixture model, for which precise asymptotics for the max-$\ell_2$ margin was established in [29].

In Section 3.5.2, we prove a universality result of the following form: the value of the max-$\ell_1$-margin remains the same (asymptotically) under two different settings where the distribution of the features entered in the boosting algorithm vary. To describe in detail, suppose the observed data $\{x_i, y_i\}$ still arises from the data-generating distribution considered for our aforementioned point-by-point contributions. However, the features feeding to the boosting algorithm (and thus in calculating the margin) are more complicated than the raw features $x_i$'s. We consider two different kinds of boosting features: (i) features $a_i$ that take the form of

a random feature model $a_i = \sigma(F^\top x_i)$ [53, 74, 78], (ii) features $b_i = \mu_0 \mathbf{1} + \mu_1 F^\top x_i + \mu_2 z_i$, where the constants $\mu_0$, $\mu_1$, $\mu_2$ are calibrated appropriately to match moments of $a_i$'s and $b_i$'s. Here, $F$ is a random matrix in $\mathbb{R}^{p\times d}$ and $z_i$ has i.i.d. $\mathcal{N}(0,1)$ entries, independent of everything else. In each case, the max-$\ell_1$-margin is calculated using the formula $\kappa_{n,\ell_1}(\{r_i, y_i\}_{1\le i\le n}) := \max_{\|\theta\|_1\le 1} \min_{1\le i\le n} y_i r_i^\top \theta$, where $r_i = a_i$ (resp., $b_i$) in Case (i) (resp., Case (ii)). Section 3.5.2 establishes that, when $p$, $d$ both scale linearly with $n$, the (scaled) max-$\ell_1$-margin has the same limiting value under both settings.

The aforementioned result holds under certain assumptions on the random feature matrix $F$ and the nonlinearity $\sigma(\cdot)$. (See Section 3.5.2 for details). But note that, conditional on $F$, $b_i$ is Gaussian whereas $a_i$ is not. This universality result suggests that the margin value is asymptotically insensitive, at least under some settings, to nuanced properties of the feature distribution. Thus, results that apply for the Gaussian case might be relevant for certain non-Gaussian feature distributions as well. We further validate this through empirical observations in Section 3.5.2. On the technical front, our universality result starts with a leave-one-out argument from [53]. However, [53] considered loss functions satisfying smoothness and strong-convexity assumptions that are violated in our setting. This leads to technical challenges that we handle by establishing new analytic results (Section 3.5.2 and Appendix D.2).

*Finite sample performance.* Our results are asymptotic in nature, and here we test their finite sample accuracy via a simple simulation. Consider a grid of values for the over-parametrization ratio $\psi \in \Psi \subset [0, 6]$, and a data-generating process where the covariates $x_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \boldsymbol{I}_p)$, and the response $y_i|x_i = +1$ with probability $\sigma(x_i^\top \theta_\star)$ where $\sigma(t) = 1/(1+e^{-t})$, and $y_i|x_i = -1$ otherwise. Each coordinate of $\theta_\star$ is drawn i.i.d. from a Gaussian $\mathcal{N}(0, 1/p)$. For each $\psi \in \Psi$, we generate multiple samples of size $n = 400$, and calculate the max-$\ell_1$-margin by two methods: (i) the numerical solution $\kappa_{n,\ell_1}$ to the corresponding linear program (LP) in (1.3); the blue points in Figure 1(a) depict these values (appropriately scaled) and (ii) the asymptotic value $\kappa_\star(\psi, \mu)$ predicted by our analytic formula in Theorem 3.1; the red points labeled as CGMT in Figure 1(a) represent these values. Calculating our theoretical predictions involves solving a complex *nonlinear system of equations* defined in (3.9). This involved computing integrals, which we approximate via Monte-Carlo sums (5000 samples). Figure 1(b) compares the corresponding out-of-sample prediction error: the blue points show the generalization error $\mathbb{P}_{\mathbf{x},\mathbf{y}}(\mathbf{y}\cdot\mathbf{x}^\top\hat{\theta}_{n,\ell_1} < 0)$, when $\hat{\theta}_{n,\ell_1}$ is calculated from the LP, whereas the red points depict the asymptotic value predicted by our theory (Theorem 3.2). In both

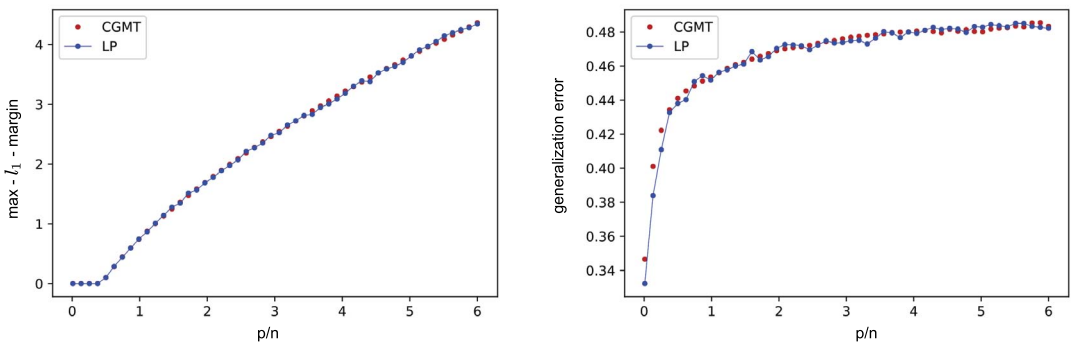


FIG. 1. *x-axis*: *Ratio* $p/n$. *y-axis*: (*a*) *Left*: *max-*$\ell_1$*-margin* (*as in equation* (3.2)), *the blue points are obtained by solving the LP in* (1.3) *and averaging its solution over* 10 *independent simulation runs. The red points are obtained by numerically evaluating the formula in RHS of* (3.2). (*b*) *Right*: *Generalization error, the blue points are obtained by calculating the generalization error of* $\hat{\theta}_{n,\ell_1}$ *that forms the solution in* $\theta$ *of the LP* (1.3), *and this is averaged over* 10 *simulation runs. The red points are obtained by numerically evaluating the formula in RHS of* (3.2).

cases, the points align well, demonstrating that our theory, albeit asymptotic, shows satisfactory finite sample accuracy. In this example, the threshold for separability was around 0.43 [24]. This is also evidenced in the plot—the max-$\ell_1$-margin is positive (resp., zero) above (resp., below) this threshold, and as expected, our theory matches the numerics accurately above the threshold.

*Organization*. The rest of the paper is organized as follows. Section 2 introduces some crucial preliminaries that are heavily used through the rest of the paper. Section 3 presents our main results, whereas a proof sketch and description of our technical contributions is presented in Section A (details are deferred to the Appendix). Section 4 discusses relevant literature that has been omitted from this Introduction. Finally, Section 5 concludes with a discussion on possible directions for future work.

**2. Formal setup and preliminaries.** This section introduces our formal setup. Unless otherwise mentioned, we consider a sequence of problems $\{y(n), X(n), \theta_\star(n)\}_{n\geq 1}$, such that $y(n) \in \mathbb{R}^n$, $\theta_\star(n) \in \mathbb{R}^{p(n)}$ and $X(n) \in \mathbb{R}^{n\times p(n)}$, where the $i$th row $x_i \sim \mathcal{N}(0, \Lambda(n))$, and the $i$th entry of $y(n)$ satisfies

$$y_i|x_i \overset{\text{i.i.d.}}{\sim} \begin{cases} +1, & \text{w.p. } f\big(\langle \theta_\star(n), x_i\rangle\big), \\ -1, & \text{w.p. } 1 - f\big(\langle \theta_\star(n), x_i\rangle\big). \end{cases} \tag{2.1}$$

Above, $\Lambda(n) \in \mathbb{R}^{p(n)\times p(n)}$ is a diagonal covariance matrix and $f$ is any nondecreasing continuous function bounded between 0 and 1. Recall that we consider the asymptotic regime (1.1), that is, $p(n)/n \to \psi \in (0, \infty)$. We require certain structural assumptions on the covariate distributions and the regression vector sequence that is described below. Conceptually, four factors determine the structure of the problem: overparametrization $\psi$, signal strength $\rho$, link function $f$ and a limiting measure $\mu$ defined in Assumption 2. Later, Section 3.5 will investigate models beyond (2.1).

ASSUMPTION 1. Let $\lambda_i(n)$ denote the eigenvalues of $\Lambda(n)$. Assume that there exists a positive constant $0 < c < 1$ such that $c \leq \lambda_i(n) \leq 1/c$, $\forall 1 \leq i \leq p(n)$ and for all $n$ and $p$.

ASSUMPTION 2. Define $\rho(n) \in \mathbb{R}$ and $\bar{w}(n) \in \mathbb{R}^{p(n)}$ such that

$$\rho(n) := \big(\theta_\star(n)^\top \Lambda(n)\theta_\star(n)\big)^{1/2} \quad \text{and} \quad \bar{w}_i(n) := \sqrt{p}\frac{\sqrt{\lambda_i(n)}\langle \theta_\star(n), e_{i,p}\rangle}{\rho(n)}, \tag{2.2}$$

where $e_{i,p}$ denotes the canonical vector in $\mathbb{R}^p$ with 1 in the $i$th entry and 0 elsewhere. Assume

$$\rho(n) \to \rho \tag{2.3}$$

with $0 < \rho < \infty$. Assume in addition that the empirical distribution of $\{(\lambda_i(n), \bar{w}_i(n))\}_{i=1}^{p(n)}$ converges to a probability distribution $\mu$ on $\mathbb{R}_{>0} \times \mathbb{R}$, in the Wasserstein-2 distance, that is,

$$\frac{1}{p}\sum_{i=1}^{p} \delta_{(\lambda_i, \bar{w}_i)} \overset{W_2}{\Rightarrow} \mu. \tag{2.4}$$

REMARK 2.1. Note that Assumption 1 and (2.3) together imply that $\sum_{j=1}^{p} \theta_\star(n)_j^2 = O(1)$. If all the entries of $\theta_\star$ are of the same order, this yields $\theta_{\star,i} = O(1/\sqrt{p})$. This also justifies why we include $\sqrt{p}$ in the numerator of $\bar{w}_i$. The convergence in $W_2$ equivalently means weak convergence and convergence of the second moments (see, for instance, [75, 103]). In particular, this implies that $\int w^2 \mu(d\lambda, dw) = 1$.

ASSUMPTION 3. Finally, assume that

$$\|\bar{w}(n)\|_\infty \le C', \quad \text{and} \quad \|\bar{w}(n)\|_1/p > C'' \tag{2.5}$$

for all $n$ and $p$, for some constants $C', C'' > 0$.

*Linear separability.* We assume that our sequence of problem instances is (asymptotically) linearly separable in the following sense:

$$\lim_{n,p(n)\to\infty} \mathbb{P}\big(\exists\theta \in \mathbb{R}^p, y_i x_i^\top \theta > 0 \text{ for } 1 \le i \le n\big) = 1. \tag{2.6}$$

For the model specified in (2.1), it turns out that (2.6) is satisfied if and only if the overparametrization ratio exceeds a phase transition threshold $\psi > \psi^\star(\rho, f)$. It is well known that the separability event is equivalent to the event that the maximum likelihood estimate is attained at infinity [2], and this has been a problem of intense study in classical statistics and information theory [28, 61, 86]. More recently, [24] derived the separability threshold $\psi^\star(\rho, f)$ for a logistic regression model (when $f$ is the sigmoid function). A similar phenomenon extends to other functions $f$ as well, as subsequently characterized by [75]. To describe this phase transition threshold, consider the following bivariate function $F_\kappa : \mathbb{R} \times \mathbb{R}_{\ge 0} \to \mathbb{R}_{\ge 0}$ defined for any $\kappa \ge 0$:

$$F_\kappa(c_1, c_2) := \big(\mathbb{E}\big[(\kappa - c_1 Y Z_1 - c_2 Z_2)_+^2\big]\big)^{\frac{1}{2}} \quad \text{where}$$

$$\begin{cases} Z_2 \perp (Y, Z_1), \\ Z_i \sim \mathcal{N}(0,1), \quad i = 1, 2, \\ \mathbb{P}(Y = +1|Z_1) = 1 - \mathbb{P}(Y = -1|Z_1) = f(\rho \cdot Z_1). \end{cases} \tag{2.7}$$

Then

$$\psi^\star(\rho, f) = \min_{c\in\mathbb{R}} F_0^2(c, 1). \tag{2.8}$$

As an example, recall that $\psi^\star(\rho, f) \approx 0.43$ in the setting of Figure 1. The above function $F_\kappa : \mathbb{R} \times \mathbb{R}_{\ge 0} \to \mathbb{R}_{\ge 0}$ will prove crucial in our subsequent theory.

*Boosting Algorithm.* For the convenience of the readers, we describe here the general *Boosting Algorithms* we work with. We begin by briefing the steps in AdaBoost [42, 43]. Suppose that each weak learner outputs a continuous decision $X_{ij} = x_i[j] \in \mathbb{R}$ and $y_i \in \{-1, +1\}$. Let $\Delta_n$ be the standard probability simplex given by $\Delta_n := \{\boldsymbol{p} \in [0,1]^n : \sum_{i=1}^n p_i = 1\}$. Suppose $Z = y \circ X \in \mathbb{R}^{n\times p}$ denotes multiplying each element in the $i$th row of $X$ by $y_i$, $i \in [n]$. At each step, AdaBoost adaptively chooses the best feature as follows:

1. Initialize: data weight $\eta_0 = 1/n \cdot \mathbf{1}_n \in \Delta_n$, parameter $\theta_0 = 0$.
2. At time $t \ge 0$:
   (a) Feature selection: $v_{t+1} := \arg\max_{v\in\{e_j\}_{j\in[p]}} |\eta_t^\top Z v|$;
   (b) Adaptive stepsize $\alpha_t$: $\alpha_t := \eta_t^\top Z v_{t+1}$;
   (c) Coordinate update: $\theta_{t+1} = \theta_t + \alpha_t \cdot v_{t+1}$;
   (d) Weight update: $\eta_{t+1}[i] \propto \eta_t[i] \exp(-\alpha_t y_i x_i^\top v_{t+1})$, normalized such that $\eta_{t+1} \in \Delta_n$.
3. Terminate after $T$ steps, and output the vector $\theta_T$.

**3. Main results.** This section will provide precise analyses of the max-$\ell_1$-margin $\kappa_{n,\ell_1}$ and the min-$\ell_1$-norm interpolant $\hat{\theta}_{n,\ell_1}$, as well as the generalization and optimization performance of *Boosting Algorithms*, in terms of the problem parameters $(\psi, \rho, \mu, f)$ introduced in Section 2.

3.1. *Max-$\ell_1$-margin and min-$\ell_1$-norm interpolant.* Recall the definition of the max-$\ell_1$-margin from (1.3). We establish that $\kappa_{n,\ell_1}$, when appropriately scaled, converges almost surely to a limit that can be explicitly characterized in terms of $\psi$, $\mu$ and $f$. To describe this limit, consider the following function first introduced in [75]: for any $(\psi, \kappa)$ pair that satisfies $\psi > \psi^{\downarrow}(\kappa)$ (see equation (3.12)), define $T : (\psi, \kappa) \to \mathbb{R}$ to be

$$T(\psi, \kappa) := \psi^{-1/2}\big[F_\kappa(c_1, c_2) - c_1 \partial_1 F_\kappa(c_1, c_2) - c_2 \partial_2 F_\kappa(c_1, c_2)\big] - s. \tag{3.1}$$

Above, $c_1 \equiv c_1(\psi, \rho, \mu, \kappa)$, $c_2 \equiv c_2(\psi, \rho, \mu, \kappa)$, $s \equiv s(\psi, \rho, \mu, \kappa)$ form the *unique* solution to the nonlinear system of equations introduced in (3.9) (Proposition 3.1 establishes uniqueness of the solution). A detailed description of this system is deferred until Section 3.2; the key point is that the system takes as input the quantities $\psi$, $\rho$, $\mu$, $\kappa$ and solves three equations in three unknowns, producing a triplet $c_1$, $c_2$, $s$. Throughout, $\mu$ and $\rho$ will be defined via (2.4) and (2.3), respectively, and if these are fixed, $c_1$, $c_2$, $s$ then simply form functions of $\psi$, $\kappa$. Note that we drop the dependence on $f$ for simplicity of the exposition; however, it is important to emphasize that $f$ enters the definition of $F_\kappa(\cdot, \cdot)$, which in turn affects the equation system.

THEOREM 3.1. *Suppose Assumptions 1–3 hold and that our sequence of problem instances obeys (2.6), that is, $\psi > \psi^\star(\rho, f)$. Then, under the asymptotic regime (1.1), the max-$\ell_1$-margin admits the limiting characterization*

$$\lim_{n\to\infty} p^{1/2} \cdot \kappa_{n,\ell_1} \overset{\text{a.s.}}{=} \kappa_\star(\psi, \rho, \mu), \tag{3.2}$$

*where*

$$\kappa_\star(\psi, \rho, \mu) = \inf\{\kappa \geq 0 : T(\psi, \kappa) = 0\}. \tag{3.3}$$

The max-$\ell_1$-margin was conjectured to be a central quantity for boosting [16]. Theorem 3.1 provides a precise high-dimensional characterization of this object under our data-generating scheme. For typical data instances, it is crucial to understand how such margin scales with the overparametrization, both theoretically and empirically, which is answered by the above theorem. This limiting result will lead to precise characterizations of statistical and computational properties of *Boosting Algorithms* in our setting, as we shall shortly see in Section 3.3. Although the result is asymptotic, the empirical margin (scaled) $\sqrt{p}\kappa_{n,\ell_1}$ agrees well with the limiting value $\kappa_\star(\psi, \rho, \mu)$, even for data sets with moderate dimensions (e.g., $n = 400$), as demonstrated by Figure 1.

Some comments regarding the limit $\kappa_\star(\psi, \rho, \mu)$ are in order. First, the limit is well defined, owing to properties of $T(\psi, \kappa)$: Section 3.2 presents an argument toward this claim. Next, (3.3) clearly demonstrates the dependence of $\kappa_\star(\psi, \rho, \mu)$ on the overparametrization ratio $\psi$. Its dependence on the signal strength $\rho$ and the distribution $\mu$ is encoded through $F_\kappa(\cdot, \cdot)$, and the parameters $c_1 \equiv c_1(\psi, \rho, \mu, \kappa)$, $c_2 \equiv c_2(\psi, \rho, \mu, \kappa)$, $s \equiv s(\psi, \rho, \mu, \kappa)$, which appear in the definition of $T(\psi, \kappa)$ (3.1).

We now proceed to study the min-$\ell_1$-norm interpolated classifier (1.2), and its precise generalization behavior in our asymptotic regime (1.1). Define

$$\mathrm{Err}_\star(\psi, \rho, \mu) = \mathbb{P}\big(c_1^\star Y Z_1 + c_2^\star Z_2 < 0\big), \tag{3.4}$$

where $c_i^\star := c_i(\psi, \rho, \mu, \kappa_\star(\psi, \rho, \mu))$, $i = 1, 2$. Together with a third parameter $s^\star \equiv s(\psi, \rho, \mu, \kappa_\star(\psi, \rho, \mu))$, $c_1^\star$, $c_2^\star$, $s^\star$ form the unique solution to the system of equations (3.9), when the inputs to the system are $\psi$, $\rho$, $\mu$ and $\kappa_\star(\psi, \rho, \mu)$, (3.2). Furthermore, $(Y, Z_1, Z_2)$ follows the joint distribution specified in (2.7); note that this depends on the problem parameters through $\rho$.

THEOREM 3.2. *Under the assumptions of Theorem 3.1, the generalization error of any min-$\ell_1$-interpolated classifier $\hat{\theta}_{n,\ell_1}$, defined in (1.2), converges almost surely to* $\mathrm{Err}_\star(\psi, \rho, \mu)$, *that is, for a new data point* $(\mathbf{x}, \mathbf{y})$ *drawn from the data-generating distribution specified in Section 2,*

$$\lim_{n\to\infty} \mathbb{P}_{(\mathbf{x},\mathbf{y})}\big(\mathbf{y}\cdot\mathbf{x}^\top\hat{\theta}_{n,\ell_1} < 0\big) \overset{\text{a.s.}}{=} \mathrm{Err}_\star(\psi, \rho, \mu). \tag{3.5}$$

Theorem 3.2 provides an exact quantification of the generalization behavior of the min-$\ell_1$-norm interpolant under our data-generating scheme. Earlier works [82, 108] already characterized the long time and infinitesimal step size limit of AdaBoost on separable data. Later, Section 3.3 will establish a slightly more refined connection between $\hat{\theta}_{n,\ell_1}$ and the AdaBoost iterates (with suitably chosen learning rates). Informally, the AdaBoost iterates arrive arbitrarily close to the min-$\ell_1$-norm interpolant, beyond a certain time threshold. Therefore, Theorem 3.2 provides two important contributions to the boosting literature, described as follows.

First, Schapire et al. [88], Breiman [16] posed a general question regarding which quantity truly governs the generalization performance of AdaBoost. Observe that in Theorem 3.2, $\mathrm{Err}_\star(\psi, \rho, \mu)$ crucially depends on $\kappa_\star(\psi, \rho, \mu)$ (3.2) through the constants $c_i^\star$. Therefore, the asymptotic max-$\ell_1$-margin precisely determines the generalization error in our setting. Since our result is asymptotically exact, Theorem 3.2 provides an answer to the question posed in [16, 88] under our assumed model. To contrast, the existing margin-based generalization upper bounds [59, 88] (that do not assume strong conditions on the data-generating distribution) scale as

$$\frac{1}{\sqrt{n}\kappa_{n,\ell_1}}\,\mathrm{Poly}(\log n) \asymp \frac{\sqrt{\psi}}{\kappa_\star(\psi, \rho, \mu)}\,\mathrm{Poly}(\log n) \gg \mathrm{Err}_\star(\psi, \rho, \mu). \tag{3.6}$$

In fact, note that the inverse of the $y$-axis in Figure 2 corresponds to the classical upper bound $(\sqrt{n}\kappa_{n,\ell_1})^{-1}$ on the generalization error, as given by equation (3.6), but this upper bound is vacuous in our setting (even overlooking the log factors) since it is worse than 0.5. As a side remark, note that Theorem 3.2 also exhibits accurate finite sample performance, as already seen in Figure 1.

Second, the constants $c_1^\star$, $c_2^\star$ carry elegant geometric and statistical interpretations. Toward establishing Theorem 3.2, it can be shown that the angle between the interpolated solution $\hat{\theta}_{n,\ell_1}$ and the target $\theta_\star$ converges in the following sense:

$$\frac{\langle\hat{\theta}_{n,\ell_1}, \theta_\star\rangle_\Lambda}{\|\hat{\theta}_{n,\ell_1}\|_\Lambda\|\theta_\star\|_\Lambda} \xrightarrow{\text{a.s.}} \frac{c_1^\star}{\sqrt{(c_1^\star)^2 + (c_2^\star)^2}}, \tag{3.7}$$

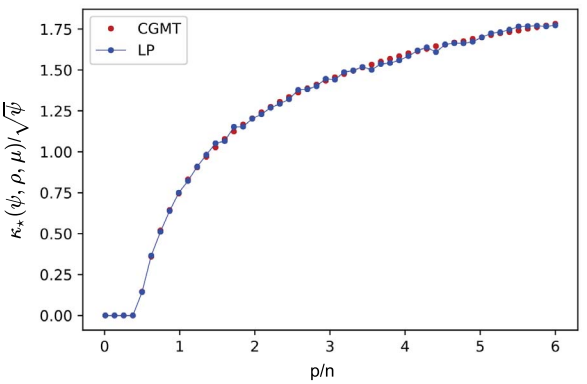


FIG. 2. *x-axis: varying ratio $\psi := p/n$. y-axis: $\kappa_\star(\psi, \rho, \mu)/\sqrt{\psi}$ (as in equation (3.6)). The setting is the same as in Figure 1. See Figure 1(a) for details on calculation of the blue and red points.*

where $\langle \theta_1, \theta_2 \rangle_\Lambda := \theta_1^\top \Lambda \theta_2$. Furthermore, $c_2^\star$ can be interpreted as the orthogonal projection, in the sense that, $\|\Pi_{(\Lambda^{1/2}\theta_\star)^\perp}(\Lambda^{1/2}\hat{\theta}_{n,\ell_1})\| \xrightarrow{\text{a.s.}} c_2^\star$.

Finally, recall the Bayes error formula, and contrast it with the test error formula (3.4) proved in Theorem 3.2,

$$\text{Err}_{\text{Bayes}}(\rho) = \mathbb{P}(YZ_1 < 0), \qquad \text{Err}_\star(\psi, \rho, \mu) = \mathbb{P}\big((c_2^\star)^{-1} c_1^\star YZ_1 + Z_2 < 0\big). \tag{3.8}$$

Then it is clear to see that $(c_2^\star)^{-1}c_1^\star$ determines how the test error of $\hat{\theta}_{n,\ell_1}$ differs from the optimal Bayes error. Therefore, Theorem 3.2 advances the literature on how the test error of boosting relates to the Bayes error [17, 56, 69, 107]: the optimality of boosting (w.r.t. the optimal Bayes classifier) is entirely determined by the magnitude of $(c_2^\star)^{-1}c_1^\star$.

The curious reader may wonder about the accuracy of our asymptotic theory for design matrices excluded from our assumptions. We investigate this sensitivity along few directions—violation of independence between the features, violation of Gaussianity of the covariates used for boosting and misspecification in the model due to missing a fraction of the relevant variables. We defer the readers to Section 3.5 for more details on these.

3.2. *The nonlinear system of equations.* We will now introduce a nonlinear system of equations that is key to the study of the max-$\ell_1$-margin and the min-$\ell_1$-norm interpolant in high dimensions, as delineated in Theorems 3.1–3.2.

DEFINITION 1. For any $\psi > 0$ and $\kappa \geq 0$, define the following system of equations in variables $(c_1, c_2, s) \in \mathbb{R}^3$,

$$\begin{aligned} c_1 &= -\mathbf{E}_{(\Lambda,W,G)\sim\mathcal{Q}}\bigg(\frac{\Lambda^{-1/2}W \cdot \mathcal{T}}{\psi^{-1/2}c_2^{-1}\partial_2 F_\kappa(c_1,c_2)}\bigg), \\ c_1^2 + c_2^2 &= \mathbf{E}_{(\Lambda,W,G)\sim\mathcal{Q}}\bigg(\frac{\Lambda^{-1/2}\mathcal{T}}{\psi^{-1/2}c_2^{-1}\partial_2 F_\kappa(c_1,c_2)}\bigg)^2, \\ 1 &= \mathbf{E}_{(\Lambda,W,G)\sim\mathcal{Q}}\bigg|\frac{\Lambda^{-1}\mathcal{T}}{\psi^{-1/2}c_2^{-1}\partial_2 F_\kappa(c_1,c_2)}\bigg|, \end{aligned} \tag{3.9}$$

where

$$\begin{aligned} \mathbf{prox}_\lambda(t) &= \arg\min_s \bigg\{\lambda|s| + \frac{1}{2}(s-t)^2\bigg\} = \text{sign}(t)\big(|t| - \lambda\big)_+, \\ \mathcal{T} &= \mathbf{prox}_s\big(\Lambda^{1/2}G + \psi^{-1/2}\big[\partial_1 F_\kappa(c_1,c_2) - c_1c_2^{-1}\partial_2 F_\kappa(c_1,c_2)\big]\Lambda^{1/2}W\big), \end{aligned} \tag{3.10}$$

and the expectation is over $(\Lambda, W, G) \sim \mu \otimes \mathcal{N}(0,1) =: \mathcal{Q}$ with $\mu$ and $F_\kappa(\cdot,\cdot)$ defined as in (2.4) and (2.7) respectively.

Note that $\Lambda$ denotes both the random variable in (3.9) and the covariance matrix in Assumption 1. Such overload of notation will prove useful in the technical derivations.

This equation system is fundamental in characterizing all of the limiting results in Section 3.1. At this point, the system may seem mysterious to the readers, but it arises rather naturally in the analysis of (1.2)–(1.3); this will be detailed in Appendix A. The max-$\ell_2$-margin has received considerable attention in the past [46, 75, 90], however, (3.9) differs significantly from the equation system considered in case of the $\ell_2$ geometry. This is natural, due to the intrinsic differences between the $\ell_2$ and $\ell_1$ geometries, and this also leads to significant technical challenges in our setting (Appendix A). Analogous systems arise in the study of high-dimensional statistical models in the proportional regime (1.1); here, the most relevant ones are the analysis of the Lasso under nonlinear measurement models [99], and that of the MLE, LRT [95, 109] and convex regularized estimators [85, 94] for logistic regression.

*Uniqueness.* Theorems 3.1–3.2 expressed our limiting results in terms of the solution to the system (3.9). It is, therefore, crucial to establish that the solution will indeed be unique. To this end, introduce the constants $\zeta$ and $\omega$ as follows:

$$\begin{aligned}\zeta &:= \big(\mathbf{E}_{(\Lambda,W)\sim\mu}\big|\Lambda^{-1/2}W\big|\big)^{-1},\\ \omega &:= \big(\mathbf{E}_{(\Lambda,W)\sim\mu}\big(W-\zeta\,\Lambda^{-1/2}\operatorname{sign}\big(\zeta\,\Lambda^{-1/2}W\big)\big)^2\big)^{1/2}.\end{aligned} \tag{3.11}$$

Define the functions $\psi_+(\kappa):\mathbb{R}_{>0}\to\mathbb{R}$, $\psi_-:\mathbb{R}_{>0}\to\mathbb{R}$ and $\psi^{\downarrow}(\kappa):\mathbb{R}_{>0}\to\mathbb{R}_{\geq 0}$ as follows:

$$\begin{aligned}\psi_+(\kappa) &= \begin{cases}0 & \text{if } \partial_1 F_\kappa(\zeta,0)>0,\\ \partial_2^2 F_\kappa(\zeta,0)-\omega^2\partial_1^2 F_\kappa(\zeta,0) & \text{if otherwise,}\end{cases}\\ \psi_-(\kappa) &= \begin{cases}0 & \text{if } \partial_1 F_\kappa(-\zeta,0)<0,\\ \partial_2^2 F_\kappa(-\zeta,0)-\omega^2\partial_1^2 F_\kappa(-\zeta,0), & \text{if otherwise,}\end{cases}\\ \psi^{\downarrow}(\kappa) &= \max\big\{\psi^\star(\rho,f),\psi_+(\kappa),\psi_-(\kappa)\big\},\end{aligned} \tag{3.12}$$

where $\psi^\star(\rho,f)$ is given by (2.8).

PROPOSITION 3.1. *For any $(\psi,\kappa)$ pair satisfying $\psi>\psi^{\downarrow}(\kappa)$, under Assumptions 1–3, the system of equations (3.9) admits a unique solution that satisfies $(c_1,c_2,s)\in\mathbb{R}\times\mathbb{R}_{>0}\times\mathbb{R}_{>0}$.*

Our proof for Proposition 3.1 adapts insights from [75] to the case of $\ell_1$ geometry, however, the definition of $\omega,\zeta$ in the threshold $\psi^{\downarrow}(\kappa)$, (3.12), differs from the $\ell_2$ case. Now, it can be shown that $F_\kappa(\cdot,\cdot)$ satisfies: (i) $(\psi,\kappa)\mapsto T(\psi,\kappa)$ is continuous on its domain, (ii) for any fixed $\kappa>0$, $T(\psi,\kappa)$ is strictly decreasing in $\psi$, (iii) for any fixed $\psi>0$, $T(\psi,\kappa)$ is strictly increasing in $\kappa$ ([75], Section B.5, Proposition 4.1). Further, using the definition of $\psi^{\downarrow}(\kappa)$, and once again properties of $F_\kappa(\cdot,\cdot)$, one can establish that $\lim_{\psi\to\infty}T(\psi,\kappa)<0$, whereas $\lim_{\psi\downarrow\psi^{\downarrow}(\kappa)}T(\psi,\kappa)>0$ and, moreover, $\lim_{\kappa\to\infty}T(\psi,\kappa)=\infty$. Putting all of these together yields that the region $\{(\psi,\kappa):\psi>\psi^{\downarrow}(\kappa)\}$ contains the region $\{(\psi,\kappa):T(\psi,\kappa)=0\}$. This ensures (3.3) is well defined, and that $c_1^\star$, $c_2^\star$, $s^\star$ are unique. We defer to the Appendix for proof of Proposition 3.1.

3.3. *Boosting in high dimensions.* We turn our attention to the *Boosting Algorithm* described in Section 2. The path of boosting iterates was studied in infinite time and infinitesimal stepsize in [82, 108]. Here, we establish a sharp analysis of the number of iterations necessary for the AdaBoost iterates to approximately maximize the $\ell_1$-margin with arbitrary accuracy.

THEOREM 3.3. *Under the assumptions of Theorem 3.1, with a suitably chosen learning rate (specified in Corollary* A.1*), the sequence of iterates $\{\hat\theta^t\}_{t\in\mathbb{N}}$ obtained from the* Boosting Algorithm *obeys the following property: for any $0<\epsilon<1$, when the number of iterations $t$ satisfies*

$$t\geq T_\epsilon(n)\quad \textit{with}\ \lim_{n\to\infty}\frac{T_\epsilon(n)}{n\log^2 n}\overset{\text{a.s.}}{=}\frac{12\psi}{\kappa_\star^2(\psi,\rho,\mu)}\epsilon^{-2}, \tag{3.13}$$

*the solution $\hat\theta^t/\|\hat\theta^t\|_1$ forms $(1-\epsilon)$-approximation to the Min-$\ell_1$-interpolated classifier, that is, almost surely,*

$$(1-\epsilon)\cdot\kappa_\star(\psi,\rho,\mu)\leq\liminf_{n\to\infty}\bigg(p^{1/2}\cdot\min_{i\in[n]}\frac{y_i x_i^\top\hat\theta^t}{\|\hat\theta^t\|_1}\bigg)$$

$$\leq \limsup_{n\to\infty}\left(p^{1/2}\cdot \min_{i\in[n]} \frac{y_i x_i^\top \hat{\theta}^t}{\|\hat{\theta}^t\|_1}\right)$$
$$\leq \kappa_\star(\psi,\rho,\mu).$$

The above result is obtained by combining our Theorem 3.1 with a careful nonasymptotic analysis of AdaBoost allowing for an explicitly-specified learning rate that builds upon existing works on margin maximization rates, rescaling and shrinkage techniques, and the mirror descent connections of AdaBoost (see [26, 38, 50, 54, 98, 108] and references cited therein). Together with Theorem 3.2, this result establishes a precise characterization of the computational and statistical behavior of AdaBoost for all iterations above the threshold $T_\epsilon(n)$, and complements the classical margin upper bounds [59, 88]. Thus, Theorem 3.3 reinforces a crucial conclusion from Section 3.1—the max-$\ell_1$-margin is the key quantity governing the generalization error of AdaBoost in our setting.

Aside from strengthening this conclusion, for separable data with a large and comparable number of samples and features, the theorem informs a stopping rule for *Boosting Algorithms* that ensures good generalization behavior. Note that, for any numerical accuracy $\epsilon$, the stopping time $T_\epsilon(n)$ has an asymptotic characterization that contributes new insights to the computational properties of AdaBoost. To see this, Figure 2 plots the scaled margin limit $\psi^{-1/2}\kappa_\star(\psi,\rho,\mu)$ as a function of $\psi$, in the setting of Figure 1. The increase in this (scaled) limit as a function of $\psi$, together with (3.13), implies that the larger the overparametrization ratio, the smaller the threshold $T_\epsilon(n)$. Therefore, *overparametrization leads to faster optimization*. Furthermore, even in terms of the optimization performance, the max-$\ell_1$-margin is once again the central quantity in our setting, as elucidated by (3.13).

REMARK 3.1. A natural question may arise at this point: does the max-$\ell_1$-margin studied here, when appropriately scaled, differ significantly from the $\ell_2$-margin [75]? Note that the rescaled $\ell_1$-margin is always larger than the $\ell_2$-margin, denoted by $\kappa_{n,\ell_2}$, since

$$\kappa_{n,\ell_2} \leq \sqrt{p}\cdot\kappa_{n,\ell_1}, \quad \text{where } \kappa_{n,\ell_2} := \max_{\|\theta\|_2\leq 1}\min_{1\leq i\leq n} y_i x_i^\top \theta. \tag{3.14}$$

A comparison of Figure 2 with [75], Figure 1, shows that the range for the $\ell_1$-margin is roughly twice that for the $\ell_2$ case, demonstrating that these behave differently, even after appropriate scaling.

*Proportion of activated features for AdaBoost.* The connection between the boosting solution and max-$\ell_1$-margin suggests that AdaBoost effectively converges to a sparse classifier. Motivated to understand the geometry of the solution better, the following theorem studies the proportion of active features when the training error vanishes along the path of AdaBoost.

COROLLARY 3.1. *Let $S_0(p)$ denote the number of features selected the first time $t$ when the* Boosting Algorithm *achieves zero training error (with an initialization of $\hat{\theta}^0 = 0$), in the sense that*

$$S_0(p) := \#\{j\in[p] : \hat{\theta}_j^t \neq 0\}, \quad \textit{where } \frac{1}{n}\sum_{i=1}^n \mathbb{I}_{y_i x_i^\top \hat{\theta}^t \leq 0} = 0. \tag{3.15}$$

*Under the assumptions of Theorem 3.3, $S_0(p)$, scaled appropriately, is asymptotically bounded by*

$$\limsup_{p\to\infty} \frac{S_0(p)}{p\log^2 p} \leq \frac{12}{\kappa_\star^2(\psi,\rho,\mu)}, \quad a.s. \tag{3.16}$$

This corollary provides specific insights into the geometry of the boosting solution, by quantifying the maximum number of coordinates that may be nonzero. Note once again that the bound involves the max-$\ell_1$-margin limit, and suggests that the larger the margin, the sparser the solution (with zero training error). Thus, our limit $\kappa_\star(\psi, \rho, \mu)$ may even be central for determining the geometric structure of the boosting solution (at least under our data-generating scheme), beyond its foregoing roles in terms of generalization and optimization. Note also that the margin grows as a function of $\psi$ (Figure 1). This further suggests that larger the overparametrization, less the number of activated coordinates for certain data-generating processes.

3.4. *A new class of Boosting Algorithms.* This section studies variants of AdaBoost that converge to the max-$\ell_q$-margin direction for general $q \geq 1$. We also characterize the generalization error and optimization performance of a class of such algorithms, through a study of the max-$\ell_q$-margin and the min-$\ell_q$-norm interpolant beyond the case of $q = 1$. This complements the study of general $\ell_q$ constraints, that was initiated by [82] (see also [50] and references therein). To this end, define the max-$\ell_q$-margin to be

$$\kappa_{n,\ell_q} := \max_{\|\theta\|_q \leq 1} \min_{1 \leq i \leq n} y_i x_i^\top \theta, \tag{3.17}$$

and the corresponding min-$\ell_q$-norm interpolant to be

$$\hat{\theta}_{n,\ell_q} \in \arg\min_{\theta} \|\theta\|_q, \quad \text{s.t. } y_i x_i^\top \theta \geq 1, 1 \leq i \leq n. \tag{3.18}$$

Denote $q_\star \geq 1$ to be the conjugate index of $q$, with $1/q_\star + 1/q = 1$, and consider the following algorithm.

*AdaBoost variant corresponding to $\ell_q$ geometry:*

1. Initialize: $\eta_0 = 1/n \cdot \mathbf{1}_n \in \Delta_n$, and parameter $\theta_0 = 0$.
2. At time $t \geq 0$:
   (a) Update direction: $v_{t+1} := \arg\max_{v \in \mathbb{R}^p, \|v\|_q = 1} \langle Z^\top \eta_t, v \rangle$;
   (b) Adaptive stepsize: $\alpha_t(\beta) = \beta \cdot \|Z^\top \eta_t\|_{q_\star}$, with $0 < \beta < 1$ being a shrinkage factor.
   (c) Parameter update: $\theta_{t+1} = \theta_t + \alpha_t \cdot v_{t+1}$;
   (d) Weight update: $\eta_{t+1}[i] \propto \eta_t[i] \exp(-\alpha_t y_i x_i^\top v_{t+1})$, normalized such that $\eta_{t+1} \in \Delta_n$.
3. Terminate after $T$ steps, and output the vector $\theta_T$.

This algorithm converges to the max-$\ell_q$-margin direction, as indicated by the following corollary.

COROLLARY 3.2 (Boosting converges to max-$\ell_q$-margin direction). *Let $q \geq 1$. Consider the aforementioned boosting algorithm with learning rate $\alpha_t(\beta) := \beta \cdot \eta_t^\top Z v_{t+1}$, where $\beta < 1$. Assume that $|X_{ij}| \leq M$ for $i \in [n]$, $j \in [p]$. Then after $T$ iterations, the boosting iterates $\theta_T$ converge to the max-$\ell_q$-margin direction in the following sense: for any $0 < \epsilon < 1$,*

$$\kappa_{n,\ell_q} \geq \min_{i \in [n]} \frac{y_i x_i^\top \theta_T}{\|\theta_T\|_q} > \kappa_{n,\ell_q} \cdot (1 - \epsilon), \tag{3.19}$$

*where $T \geq \log(1.01ne) \cdot \frac{2p^{\frac{2}{q_\star}} M^2 \epsilon^{-2}}{\kappa_{n,\ell_q}^2}$. The shrinkage factor is chosen as $\beta = \frac{\epsilon}{p^{\frac{2}{q_\star}} M^2}$.*

Utilizing arguments similar to that for Theorems 3.1–3.2, it can be shown that the max-$\ell_q$-margin and the corresponding min-$\ell_q$-norm interpolant admit analogous characterizations with a system of equations that differs from (3.9), all else remaining the same. To introduce the equation system corresponding to general $\ell_q$ geometry, define the proximal mapping operator of the function $f_\lambda(t) = \lambda|t|^q$, for $\lambda > 0$, $q \geq 1$, to be

$$\mathbf{prox}_\lambda^{(q)}(t) := \arg\min_s \left\{ \lambda|s|^q + \frac{1}{2}(s-t)^2 \right\}. \tag{3.20}$$

With

$$t^\star := -\frac{\Lambda^{-1/2} G + \psi^{-1/2}[\partial_1 F_\kappa(c_1, c_2) - c_1 c_2^{-1} \partial_2 F_\kappa(c_1, c_2)]\Lambda^{-1/2} W}{\psi^{-1/2} c_2^{-1} \partial_2 F_\kappa(c_1, c_2)},$$

$$\lambda^\star := \frac{\Lambda^{-1} s}{\psi^{-1/2} c_2^{-1} \partial_2 F_\kappa(c_1, c_2)}$$

define

$$h^\star = \mathbf{prox}_{\lambda^\star}^{(q)}(t^\star).$$

Consider the system of equations

$$c_1 = \langle \Lambda^{1/2} h^\star, W \rangle_{L_2(\mathcal{Q})}, \qquad c_1^2 + c_2^2 = \|\Lambda^{1/2} h^\star\|_{L_2(\mathcal{Q})}^2, \qquad \|h^\star\|_{L_q(\mathcal{Q})} = 1, \tag{3.21}$$

where, recall from Definition 1 that $Q = \mu \times \mathcal{N}(0,1)$. It is not hard to see that this system reduces to (1) for $q = 1$.

COROLLARY 3.3. *Under the assumptions of Theorem 3.1 and for $1 \leq q \leq 2$, the max-$\ell_q$-margin obeys*

$$p^{\frac{1}{q}-\frac{1}{2}} \kappa_{n,\ell_q} \xrightarrow{\text{a.s.}} \kappa_\star^{(q)}(\psi, \rho, \mu), \tag{3.22}$$

*where $\kappa_\star^{(q)}(\psi, \rho, \mu)$ satisfies (3.3), with $T(\psi, \kappa)$ of the same form as in (3.1), but with $c_1$, $c_2$, $s$ given by the solution to (3.21). Simultaneously, the generalization error of the min-$\ell_q$-norm interpolant can be characterized using (3.5), but when $c_1^\star$, $c_2^\star$, $s^\star$ is replaced by the solution to (3.21), when $\kappa_\star^{(q)}(\psi, \rho, \mu)$ is input instead of $\kappa_\star(\psi, \rho, \mu)$.*

Corollary 3.2 then establishes that all properties of AdaBoost presented in Section 3.3 continue to hold (after appropriate scalings) for the generalized versions of AdaBoost considered here for $1 \leq q \leq 2$, with (3.9) swapped for (3.21). Once again, observe that the max-$\ell_q$-margin is crucial for understanding properties of these variants of AdaBoost. In terms of proofs, our technical contributions in the context of the max-$\ell_1$-margin are sufficiently general, and can be adapted to establish the results in this section. Extensions to the case of $q > 2$ may be feasible if one imposes a condition stronger than convergence in $W_2$ (in Assumption 2).

The curious reader might wonder how the interpolant behavior changes as a function of $q$. Recall (Section 3) that the term $(c_2^\star)^{-1} c_1^\star$ governs the difference between the generalization error and the optimal Bayes error. Characterizing the minimizer of this quantity requires a refined analysis of the system of equations (3.21). Analogous calculations have been carried out elsewhere, for example, for choosing optimal loss functions for inference [1, 7, 97] and optimal bridge penalty [104] for variable selection. Our initial simulations suggest that, for our problem, there may not be a universally optimal choice of $q$. Instead, the optimal value may depend on subtle properties of the data-generating process. We defer further investigations along this line to future work.

REMARK 3.2. Note that Corollary 3.3 assumes the data is asymptotically linearly separable, that is, $\psi > \psi^\star(\rho, f)$. This separability threshold is an inherent property of the sequence of problem instances, and does not depend on the geometry under which the max-margin is considered in (3.22).

3.5. *Robustness to assumptions*. The theory presented so far provides precise characterizations of the $\ell_1$ margin, interpolant and in turn AdaBoost, but relies, nonetheless, upon assumptions on the data generating process (2.1). This section explores relaxations of these assumptions along a few natural directions: (a) going beyond the assumption of independence between the covariates, (b) analyzing sensitivity to the Gaussianity assumption, (c) understanding implications of certain model misspecification. For the latter, we explore a common source of misspecification that occurs when the model misses a fraction of relevant variables. Studying AdaBoost and the max-$\ell_1$-margin under such varied settings, we will uncover that the general insights underlying our proposed theory persist across the board, suggesting the possibility of extending our analyses to a broader class of data generation schemes.

3.5.1. *Beyond independent covariates*. This section will focus on data-generation schemes with dependent covariates. Our exact asymptotics continue to hold for a class of such design matrices. We present results in the context of two models in an increasing order of complexity the first (resp., second) involves a feature covariance matrix that is a rank-one (resp., rank-two) perturbation of a diagonal. Extensions to rank-$\ell$ perturbations are feasible (see Appendix D.1). The reader should take this section as a proof of concept that our results can be extended to dependent covariates in certain settings.

As a first step toward understanding dependent covariates, consider a simple Gaussian mixture model:

$$\mathbb{P}(y_i = +1) = 1 - \mathbb{P}(y_i = -1) = \upsilon \in (0, 1), \tag{3.23}$$

$$x_i | y_i \sim \mathcal{N}(y_i \cdot \theta_\star, \Lambda), \tag{3.24}$$

where $\Lambda \in \mathbb{R}^{p\times p}$ is a diagonal matrix. By the Bayes formula, the conditional distribution of $y_i|x_i$ can be captured through a logistic model, with $\mathbb{P}(y_i = +1|x_i) = f(\log \frac{\upsilon}{1-\upsilon} + \langle \Lambda^{-1}\theta_\star, x_i\rangle)$ and $f(t) = 1/(1+e^{-t})$. The covariate distribution obeys a mixture of Gaussians but the marginal covariance is given by $\mathrm{Cov}(x_i) = 4\upsilon(1-\upsilon)\theta_\star\theta_\star^\top + \Lambda$ (thus called the spiked covariance model). Compared to the diagonal covariance as in (2.1), the setting considered here therefore goes beyond independent covariates by introducing a rank-one spike to the diagonal covariance $\Lambda$.

Similar to Assumption 2, let $p(n)/n = \psi$ and denote

$$\frac{1}{p}\sum_{i=1}^{p} \delta_{(\lambda_i, \sqrt{p}\theta_\star^\top e_i)} \overset{W_2}{\Rightarrow} \mu. \tag{3.25}$$

Define a new function $\bar{F}_\kappa : \mathbb{R}\times\mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ with parameter $\kappa \geq 0$,

$$\bar{F}_\kappa(c_1, c_2) := \big(\mathbb{E}\big[(\kappa - c_1 - c_2 Z)_+^2\big]\big)^{\frac{1}{2}} \quad \text{where } Z \sim \mathcal{N}(0,1). \tag{3.26}$$

Denote a triplet of random variables $(\Lambda, \Theta, G) \sim \mu \otimes \mathcal{N}(0,1) =: \mathcal{Q}$ with $\mu$ given by (3.25), and for any $\psi > 0$, define the following system of equations in variables $(c_1, c_2, s) \in \mathbb{R}^3$:

$$\begin{aligned} c_1 &= -\mathbf{E}_{(\Lambda,\Theta,G)\sim\mathcal{Q}}\Big(\frac{\Lambda^{-1}\Theta \cdot \mathbf{prox}_s(\Lambda^{1/2}G + \psi^{-1/2}\partial_1\bar{F}_\kappa(c_1,c_2)\Theta)}{\psi^{-1/2}c_2^{-1}\partial_2 F_\kappa(c_1,c_2)}\Big), \\ c_2^2 &= \mathbf{E}_{(\Lambda,\Theta,G)\sim\mathcal{Q}}\Big(\frac{\Lambda^{-1/2}\,\mathbf{prox}_s(\Lambda^{1/2}G + \psi^{-1/2}\partial_1\bar{F}_\kappa(c_1,c_2)\Theta)}{\psi^{-1/2}c_2^{-1}\partial_2\bar{F}_\kappa(c_1,c_2)}\Big)^2, \end{aligned} \tag{3.27}$$

$$1 = \mathbf{E}_{(\Lambda,\Theta,G)\sim\mathcal{Q}}\left|\frac{\Lambda^{-1}\,\mathbf{prox}_s(\Lambda^{1/2}G + \psi^{-1/2}\partial_1\bar{F}_\kappa(c_1,c_2)\Theta)}{\psi^{-1/2}c_2^{-1}\partial_2\bar{F}_\kappa(c_1,c_2)}\right|.$$

Then, in the regime where the data is asymptotically linearly separable (see [29], Proposition 3.1, for the linear separability threshold for this problem), the max-$\ell_1$-margin and min-$\ell_1$-norm interpolant obey the limiting characterizations from Theorems 3.1–3.2, with the system of equations given by (3.27), and $F_\kappa(c_1,c_2)$ substituted by (3.26). Note that [29] analyzed the max-$\ell_2$-margin for a (misspecified) logistic model and the Gaussian mixture model (3.23)–(3.24) through a unified CGMT based analysis. Due to crucial differences between $\ell_1$ and $\ell_2$ geometries, the $\ell_1$ case, (or for that matter, any $\ell_q$ with $q \neq 2$) does not follow directly from these results. We will elaborate on this point in Appendix A.

We can further this characterization to analogous settings where the marginal covariance between features contains a finite rank perturbation of a diagonal matrix. To provide a precise description, consider an extension of (3.23)–(3.24), where (3.23) remains the same but (3.24) changes to

$$x_i = y_i\theta_\star + m_i\tilde{\theta} + \tilde{x}_i, \tag{3.28}$$

with $(y_i, m_i, \tilde{x}_i)$ independent of each other, $m_i$ any random variable symmetric around zero, $\tilde{x}_i \sim \mathcal{N}(0,\Lambda)$ with $\Lambda$ diagonal. The observed data contains only $(y_i, x_i)$, and thus, the $m_i$'s may be thought of as latent random variables. Note in this case, $\mathrm{Cov}(x_i) = 4v(1-v)\theta_\star\theta_\star^\top + \mathrm{Var}(m_i)\tilde{\theta}\tilde{\theta}^\top + \Lambda$, a rank-two perturbation of a diagonal covariance matrix. The aforementioned characterization can be naturally extended with appropriate analogues of (3.25)–(3.27). We assume that the Wasserstein-2 limit of the empirical distribution sequence $\sum_{j=1}^p \delta_{(\lambda_i,\sqrt{p}\theta_\star^\top e_i,\sqrt{p}\tilde{\theta}^\top e_i)}/p$ exists, denote it by $\tilde{\mu}$, and let $(\Lambda, h_\star, \tilde{h}, G) \sim \tilde{Q} = \tilde{\mu}\otimes\mathcal{N}(0,1)$. Define the following analogue of (3.26):

$$\tilde{F}_\kappa(c_1,c_2,c_3) = \sqrt{\mathbb{E}\big[(\kappa - c_1 - c_2\tilde{Z} - c_3 M)_+^2\big]}, \tag{3.29}$$

where $M \overset{\mathrm{d}}{=} m_i$, $\tilde{Z}\sim\mathcal{N}(0,1)$, independent of $M$. Then, our Theorems 3.1–3.2 once again characterize the max-$\ell_1$-margin and min-$\ell_1$-norm interpolant behavior (see Appendix D.1 for further details) on substituting $F_\kappa(c_1,c_2)$ for $\tilde{F}_\kappa(c_1,c_2,c_3)$ and (3.9) for the following system of four equations:

$$\begin{aligned}
&c_1 = \mathbb{E}_{(\Lambda,h_\star,\tilde{h},G)\sim\tilde{Q}}[h_\star h_{\mathrm{sol}}], \qquad c_2^2 = \mathbb{E}_{(\Lambda,h_\star,\tilde{h},G)\sim\tilde{Q}}\big[(\Lambda^{1/2}h_{\mathrm{sol}})^2\big],\\
&c_3 = \mathbb{E}_{(\Lambda,h_\star,\tilde{h},G)\sim\tilde{Q}}[\tilde{h}h_{\mathrm{sol}}], \qquad 1 = \mathbb{E}_{(\Lambda,h_\star,\tilde{h},G)\sim\tilde{Q}}\big[|h_{\mathrm{sol}}|\big], \quad \text{where}\\
&h_{\mathrm{sol}} = -\frac{\mathrm{prox}_s(\Lambda^{1/2}G + \psi^{-1/2}(\partial_1\tilde{F}_\kappa(c_1,c_2,c_3)h_\star + \partial_3\tilde{F}_\kappa(c_1,c_2,c_3)\tilde{h}))}{\Lambda\psi^{-1/2}c_2^{-1}\partial_2\tilde{F}_\kappa(c_1,c_2,c_3)}.
\end{aligned} \tag{3.30}$$

Conceptually, adding an extra spike to $\mathrm{Cov}(x_i)$ increases the complexity of the equation system by introducing a new variable $c_3$. We will observe a similar phenomenon if we were to look at more complicated analogues of (3.28) with a higher rank perturbation. In general, a rank-$\ell$ perturbation leads to an $(\ell+2)$-dimensional equation system analogous to (3.30). Due to space constraints, we defer the general treatment to Appendix D.1.

For both the aforementioned models, the boosting algorithm satisfies Theorem 3.3 with the respective limiting characterization of the max-$\ell_1$-margin. A common theme across these settings is that the behavior of the margin and interpolant can be accurately characterized by a fixed-point equation system, the solution to which possesses precise physical meanings (see (3.7) and the discussion thereafter). The form of the systems vary from one model to another; however, principles underlying its origin and key proof steps remain essentially the

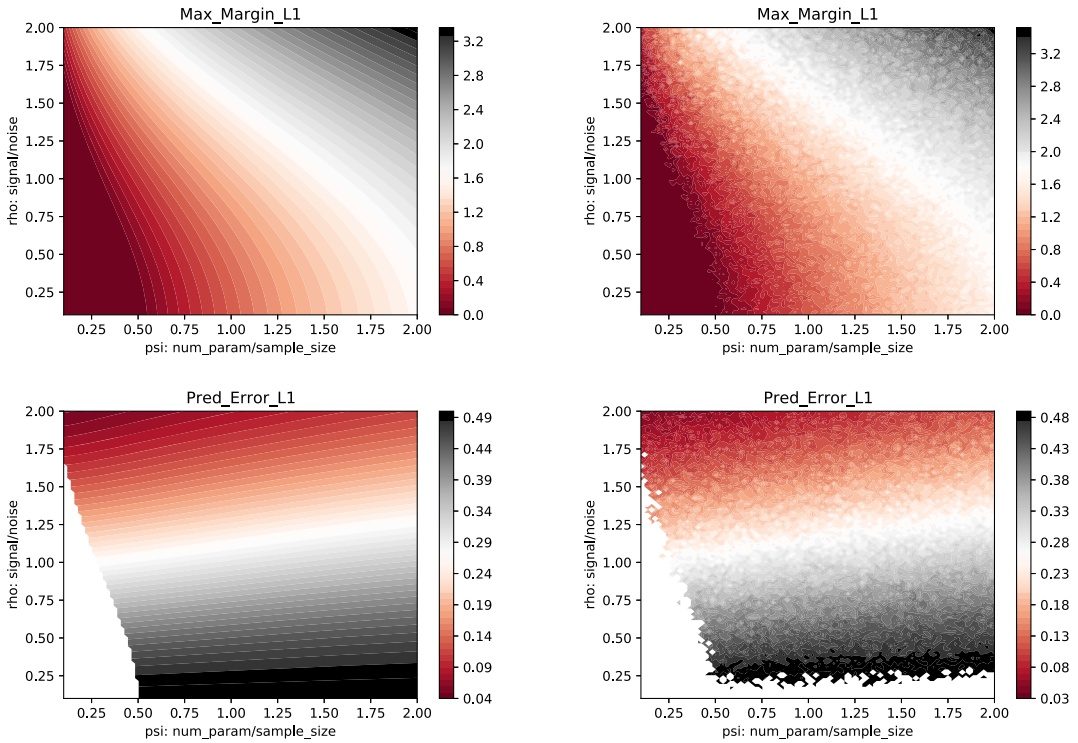


FIG. 3. *x-axis*: *Ratio* $p/n$, *y-axis*: *Signal-to-noise ratio* $\rho = (\|\sqrt{p}\theta_\star\|^2/\operatorname{Tr}(\Lambda))^{1/2}$. *The top row shows max-*$\ell_1$*-margin and bottom row the prediction error of the corresponding interpolant. The left panel plots the limits of these objects, as characterized by our asymptotic theory, while the right panel shows the corresponding finite sample values obtained by solving* (1.3) *using linear programming* (*averaged over two independent simulation runs to reduce noise*).

same (Appendix A). Once again, this is the power of our theoretical analysis in the $\ell_1$ case: we introduce a new uniform deviation argument with sufficient generality so that our proof can be adapted across several modeling schemes, as illustrated through this section.

To conclude this section, we showcase the numerical accuracy of our results for the rank-one spike case (3.23)–(3.24). The example is illustrated in Figures 3–4. Here, $\Lambda$ is always taken to be the identity matrix. The x-axis denotes the overparametrization ratio $\psi = p(n)/n$, y-axis the signal-to-noise ratio $\rho = (\|\sqrt{p}\theta_\star\|^2/\operatorname{Tr}(\Lambda))^{1/2}$ and the color encodes the value of the max-$\ell_1$-margin (top row) or prediction error of the corresponding min-$\ell_1$-norm interpolant (bottom row), respectively. Thus, for each value on the $y$-axis, we choose a different signal $\theta_\star$ so that the signal-to-noise ratio matches the given value of $\rho$. The left panel numerically solves the fixed-point equation (3.27) and presents the limits of the margin and prediction error from Theorems 3.1–3.2, obtained upon replacing (3.9) for the equation system in this rank-one spike case, (3.27). The right panel presents the max-$\ell_1$-margin in finite samples, obtained by solving the LP (1.3), along with the corresponding prediction error, and these are averaged over two independent simulation runs. As Figure 3 illustrates the finite-sample results conform to our asymptotic characterization remarkably well. Figure 4 plots the excess error, and we observe the following phenomenon: if $\psi$ stays fixed and $\rho$ increases, the excess error first increases and then decreases. Such a phenomenon indicates a hardest signal strength $\rho$ for any fixed $\psi$, where the excess error is maximized. We defer further extensions to general feature covariance matrices not covered here or in Appendix D.1 for future work. Note that the $\ell_1$ and $\ell_2$ max-margin problems differ significantly in terms of the classes of feature covariances that yield neat asymptotic limiting expressions. Appendix B explains this difference in detail.

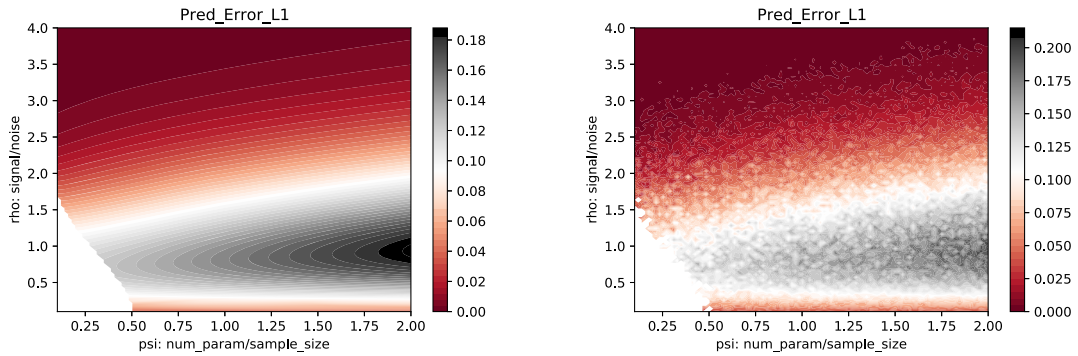


FIG. 4. *x-axis*: *Ratio* $p/n$, *y-axis*: *Signal-to-noise ratio* $\rho = (\|\sqrt{p}\theta_\star\|^2/\mathrm{Tr}(\Lambda))^{1/2}$. *The values plotted in colored-scale are the excess errors, namely* $\mathbb{P}(\mathbf{y}\mathbf{x}^\top\hat{\theta}_{n,\ell_1} < 0) - \mathbb{P}(\mathbf{y}\mathbf{x}^\top\theta_\star < 0)$. *The left panel plots the limits of these objects, as characterized by our asymptotic theory, while the right panel shows the corresponding finite sample values obtained by solving* (1.3) *using linear programming* (*averaged over two independent simulation runs to reduce noise*). *In this specific setting, if* $\psi$ *stays fixed and* $\rho$ *increases, an interesting nonmonotonic phenomenon occurs, as the excess error first increases and then decreases.*

3.5.2. *Beyond Gaussian covariates.* This section investigates the universality of the max-$\ell_1$-margin when the boosting covariates are nonlinear random features. These models are widely used in machine learning practice due to its connection to one-hidden-layer neural networks. To make the presentation clear, let us distinguish two concepts: the observed covariate-response pair $(x_i, y_i)$, and the boosting covariate-response pair $(a_i, y_i)$. To this end, consider the covariate-response pair $\{a_i \in \mathbb{R}^d, y_i\}_{i=1}^n$ fed into the boosting algorithm as stated in (2) (with the substitution $Z := [y_1 a_1, \ldots, y_n a_n]^\top \in \mathbb{R}^{n\times d}$ therein). Here, we take these "actual covariates for boosting" to be of the form $a_i = \sigma(F^\top x_i)$, with a nonlinear activation function $\sigma(\cdot)$ applied entrywise, and a random weight matrix $F \in \mathbb{R}^{p\times d}$ sampled independent of the observed $x_i$'s; thus, we call this random features. Note due to the nonlinearity of $\sigma$, the boosting features $a_i$'s are non-Gaussian even when $x_i$'s are Gaussian.

This section will show that the max-$\ell_1$-margin for the above nonlinear random features model, in the asymptotic sense, equals that of an analogous Gaussian features model, conditioned on $F$. To be concrete, we show the asymptotic equivalence of the max-$\ell_1$-margin for two models: (i) random features $a_i = \sigma(F^\top x_i) \in \mathbb{R}^d$, and (ii) analogous Gaussian features $b_i = \mu_0\mathbf{1} + \mu_1 F^\top x_i + \mu_2 z_i \in \mathbb{R}^d$, where $z_i \sim \mathcal{N}(0, I_d)$, $\mu_0 = \mathbb{E}[\sigma(Z)]$, $\mu_1 = \mathbb{E}[Z\sigma(Z)]$, $\mu_2 = \sqrt{\mathbb{E}(\sigma^2(Z)) - \mu_0^2 - \mu_1^2}$, with $Z \sim \mathcal{N}(0,1)$ independent of everything else. Here, $\mu_0, \mu_1$ are the top two Hermite coefficients of $\sigma(\cdot)$, and $\mu_2$ is the $\ell_2$ norm of the remaining Hermite coefficients. The max-$\ell_1$-margin under each model is calculated using $\kappa_{n,\ell_1}(\{r_i, y_i\}_{1\le i\le n}) := \max_{\|\theta\|_1\le 1} \min_{1\le i\le n} y_i r_i^\top\theta$, where $r_i$ equals $a_i$ or $b_i$ depending on the model. We establish that the asymptotic value of the margin (scaled by $\sqrt{p}$) remains the same irrespective of the choice of the features included in the calculation.

To formalize this result, we consider a sequence of problem instances $\{y(n), X(n), \theta^\star(n)\}_{n\ge 1}$ satisfying the conditions in Section 2, and in addition consider feature matrices $A(n)$, $B(n)$ with the $i$th row of $A(n)$ (resp., $B(n)$) given by $a_i$ (resp., $b_i$) described above. The sequence of random feature matrices $F(n)$ in the definition of $A(n)$ are taken to be of the form $F(n) = [f_1, \ldots, f_{d(n)}]$, where $f_i \sim \mathcal{N}(0, \boldsymbol{I}_p/p)$, and both $p(n), d(n)$ scale linearly with $n$. In the sequel, we suppress the dependence on $n$, whenever clear from context.

THEOREM 3.4. *Under the aforementioned conditions, if the nonlinear function* $\sigma(\cdot)$ *is odd, compactly supported, and has bounded first, second and third derivatives, then the* (*rescaled*) *max-*$\ell_1$*-margin under both fitting procedures* (*i*) *and* (*ii*) *admit the same limit in*

*probability, that is,*

$$p^{1/2} \cdot \kappa_{n,\ell_1}(\{a_i, y_i\}_{1\le i\le n}) - p^{1/2} \cdot \kappa_{n,\ell_1}(\{b_i, y_i\}_{1\le i\le n}) \xrightarrow{\mathbb{P}} 0. \tag{3.31}$$

The above theorem asserts that, asymptotically, both the nonlinear feature matrix $A(n)$ and its Gaussian counterpart $B(n)$ yield the same margin value. We next provide a brief outline of the proof. In Appendix A.1, we mention that studying the limiting value of the margin is equivalent to studying whether $\xi^{(n,p)}_{\psi,\kappa}(R) = \min_{\|\theta\|_1\le\sqrt{p}} \frac{1}{\sqrt{p}}\|(\kappa\mathbf{1} - (y\odot R)\theta)_+\|_2$ is strictly positive or not, where $R$ denotes the feature matrix used in the margin definition. This is equivalent to studying $\{\xi^{(n,p)}_{\psi,\kappa}(R)\}^2 = \min_{\|\tilde\theta\|_1\le p} \frac{1}{p}\sum_{i=1}^n(\kappa - \frac{1}{\sqrt{p}} y_i r_i^\top\tilde\theta)_+^2$, where we apply the change of variable $\tilde\theta = \sqrt{p}\theta$. Denote the Lagrange form for this problem with multiplier $\lambda$ to be $\Phi_n(R,\lambda)$. We claim that to show (3.31), it suffices to show that for all $\lambda$

$$\Phi_n(A,\lambda) - \Phi_n(B,\lambda) \xrightarrow{\mathbb{P}} 0, \tag{3.32}$$

where $A$, $B$ are the feature matrices defined under the fitting procedures (i) and (ii), respectively. To see this, denote $\lambda_A$ to be the solution to the optimization problem

$$\frac{1}{p}\min_{\|\tilde\theta\|_1\le p}\sup_{\lambda\ge 0}\sum_{i=1}^n\Big(\kappa - \frac{1}{\sqrt{p}} y_i a_i^\top\tilde\theta\Big)_+^2 + \lambda\sum_{j=1}^p(|\tilde\theta_j| - 1). \tag{3.33}$$

Then, by duality of convex programs, we have that $\{\xi^{(n,p)}_{\psi,\kappa}(A)\}^2 = \Phi_n(A,\lambda_A)$. Furthermore, $\Phi_n(B,\lambda_A) \le \frac{1}{p}\min_{\|\tilde\theta\|_1\le p}\sum_{i=1}^n(\kappa - \frac{1}{\sqrt{p}} y_i b_i^\top\tilde\theta)_+^2 + \lambda_A\sum_{j=1}^p(|\tilde\theta_j| - 1) \le \{\xi^{(n,p)}_{\psi,\kappa}(B)\}^2$. So far we have proved $\{\xi^{(n,p)}_{\psi,\kappa}(A)\}^2 \le \{\xi^{(n,p)}_{\psi,\kappa}(B)\}^2 + o_{\mathbb{P}}(1)$. Analogously, denoting $\lambda_B$ to be the solution to the optimization problem in (3.33) with $a_i$ replaced by $b_i$, and applying (3.32) with $\lambda = \lambda_B$, we obtain that $\{\xi^{(n,p)}_{\psi,\kappa}(A)\}^2 - \{\xi^{(n,p)}_{\psi,\kappa}(B)\}^2 \xrightarrow{\mathbb{P}} 0$.

To prove (3.32), we start with a leave-one-out argument adapted from [53], which in turn builds upon [35]. In [53], the authors prove that the training and generalization errors are asymptotically equivalent in a random features model and a corresponding linearized model, where the covariates have matching moments and are Gaussian conditional on the random features. However, [53] defined the training error to be based on the objective function of a penalized empirical risk minimization problem, where the loss admits derivatives up to the third order and the regularizer is strongly convex. In our setting, neither of these properties hold, and this leads to several technical challenges. To handle these, we use a specific smoothing argument and develop certain new analytic results (Appendix D.2).

To supplement our universality result, Theorem 3.4, we empirically check universality of our result across different covariate distributions used for the data-generation process. Note that this is different from the premise of Theorem 3.4. For that theorem, we considered the same data-generating distribution but different feature distribution for the covariates used in boosting, and established universality of the (asymptotic) max-$\ell_1$-margin across these settings. Now, we consider the setting of Figure 1, where the data is generated using a logistic model, and calculate the max-$\ell_1$-margin based on the linear program (1.3) (left subfigure), as well as difference between the test error and Bayes error (right subfigure), under two different settings shown in Figure 5. In the setting titled "Rademacher," each entry of the observed design is taken to be $\pm 1$ with probability $1/2$, independently of each other. In the setting titled "Gaussian," the corresponding entries are i.i.d. draws from a Gaussian distribution with first and second moments matching that of the Rademacher. In both cases, the margin values from the linear program are averaged over 10 independent runs. Observe the close match between the two settings, suggesting the applicability of our theory for a broader class of covariate distributions, beyond our theoretical results.

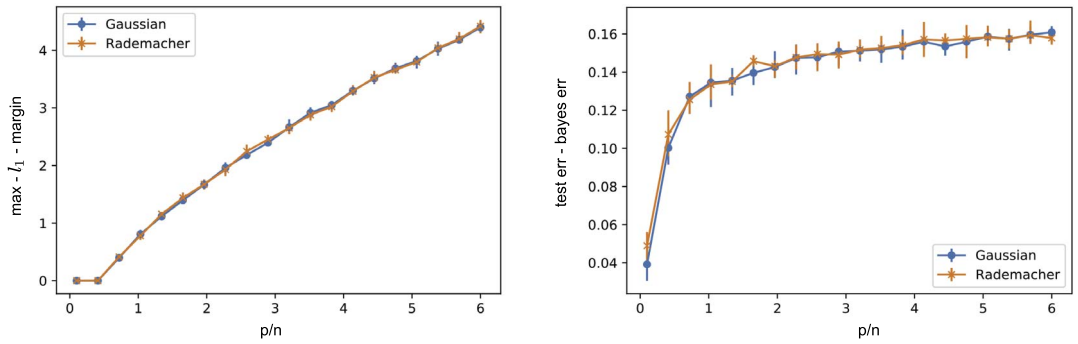


FIG. 5. *x-axis*: *Ratio* $p/n$. *y-axis*: (*Left subfigure*) *max-*$\ell_1$*-margin*, (*Right subfigure*) *Test error minus the Bayes error. The figure has the same setting as in Figure* 1, *except the covariate distribution. Here, the observed design matrix has i.i.d. entries drawn either from a Rademacher distribution or a Gaussian with matching first and second moments. The figure demonstrates universality of the margin value and the test error across these settings.*

3.5.3. *Model misspecification*. Consider the following data generating process: denote $\tilde{x}_i = (x_i^\top, z_i^\top)^\top$ where $x_i \in \mathbb{R}^p$ and $z_i \in \mathbb{R}^q$, with $x_i \sim \mathcal{N}(0, \Lambda_x)$ and $z_i \sim \mathcal{N}(0, \Sigma_z)$ independent Gaussian vectors. Here, we assume that $\Lambda_x$ is a diagonal matrix. Suppose that $y$ arises from the following conditional distribution:

$$\mathbb{P}(y_i = +1|\tilde{x}_i) = f(\tilde{x}_i^\top \theta_\star), \quad \text{with } \theta_\star := (\theta_{x,\star}^\top, \theta_{z,\star}^\top)^\top. \tag{3.34}$$

The observed data contains $n$ i.i.d. samples ($x_i \in \mathbb{R}^p$, $y_i \in \mathbb{R}$), $1 \le i \le n$, that is, only a part of the features $\tilde{x}_i$ that generate $y_i$ are included. Assume that both the seen and unseen components of the features have dimension that is large and comparable to the sample size. To model this, we assume that

$$p(n)/n = \psi > 0, \qquad q(n)/n = \phi > 0.$$

Consider that both components of $\theta_\star$, (3.34), contribute a nontrivial signal strength, in the sense that

$$\lim_{n\to\infty} (\theta_{x,\star}^\top \Lambda_x \theta_{x,\star})^{1/2} = \rho, \qquad \lim_{n\to\infty} (\theta_{z,\star}^\top \Sigma_z \theta_{z,\star})^{1/2} = \gamma,$$

where $0 < \rho, \gamma < \infty$. For any $\kappa \ge 0$, define a new function $\tilde{F}_\kappa : \mathbb{R} \times \mathbb{R}_{\ge 0} \to \mathbb{R}_{\ge 0}$,

$$\tilde{F}_\kappa(c_1, c_2) := \big(\mathbb{E}\big[(\kappa - c_1 Y Z_1 - c_2 Z_3)_+^2\big]\big)^{\frac{1}{2}},$$

$$\text{where} \quad \begin{cases} Z_3 \perp (Y, Z_1, Z_2), \\ Z_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1), \quad i = 1, 2, 3, \\ \mathbb{P}(Y = +1|Z_1, Z_2) = 1 - \mathbb{P}(Y = -1|Z_1, Z_2) = f(\rho \cdot Z_1 + \gamma \cdot Z_2). \end{cases} \tag{3.35}$$

Consider the regime where the observed data is asymptotically linearly separable, that is, $\psi + \phi$ lies above the separability threshold for this problem. We do not describe the threshold here in detail, the interested reader may find its characterization in [29], Proposition 3.1. Then the max-$\ell_1$-margin and min-$\ell_1$-norm interpolant, computed using the observed data $\{(x_i, y_i)\}_{i=1}^n$ obey the same limiting characterizations as in Theorems 3.1–3.2, with the system of equations remaining the same as in (3.9), but with $F_\kappa(c_1, c_2)$ substituted by the new function (3.35). Thus, the form of the equation system (3.9) once again remains unchanged, once we pin down the right analogue of $F_\kappa(c_1, c_2)$ in this new setting.

**4. Related literature.** This section discusses prior literature that is relevant to our problem, but were omitted from Section 1.

*Boosting*. Since its introduction in [42, 43], there has been a vast and expansive literature on boosting. [14] studied bias and variance of general arcing classifiers. A wonderful survey of early works on generalization performance of boosting, and comparisons to the optimal Bayes error can be found in [55]. Margin-based analyses were furthered in [60, 79–81]. For analysis of Boosting Algorithms based on smooth margin functions, see [84] and the references cited therein. Consistency properties were extensively studied in [12, 70–72]. Aside AdaBoost, several variants of boosting emerged over the years, accompanied by many other perspectives. Boosting for two class classifications may be viewed as additive modeling on the logistic scale [44]. Subsequently, [45] developed a general gradient boosting framework. The rate of convergence of regularized boosting classifiers was explored in [13], where the authors uncovered that some versions of boosting work especially well in high-dimensional logistic additive models. $\ell_2$-boosting, sparse boosting, twin boosting and their properties in high dimensions were extensively studied in [18–22]. We remark that our setting is different in nature from this high-dimensional Boosting literature, where a notion of sparsity (often in $\ell_1$ geometry) is typically assumed on the unknown parameter $\theta_\star$. On the contrary, the $\ell_1$ connection arises naturally in our setting, due to the nature of the AdaBoost/boosting algorithm. The rate of convergence of AdaBoost to the minimum of the exponential loss was investigated in [76]. Robust versions of boosting were proposed and extensively explored in [62]. In recent times, [39] developed novel insights into boosting, by connecting classic Boosting Algorithms for linear regression to subgradient optimization and its siblings, which might be more amenable to mathematical analysis in several settings.

*Convex Gaussian minmax theorem*. The convex Gaussian min-max theorem is a generalized and tight version of the classical Gaussian comparison inequalities [47, 48], and is obtained by extending Gordon's inequalities with the presence of convexity. The idea of merging these seemingly disparate threads dates back to [91–93], where it was used to analyze the performance of the constrained LASSO in high signal-to-noise ratio regimes. The seminal works [100–102] built and significantly extended on this idea to arrive at the CGMT, which was extremely useful for studying mean-squared errors of regularized M-estimators in high-dimensional linear models. As discussed earlier, [75] studied the asymptotic properties of the max-$\ell_2$-margin in binary classification settings, building upon CGMT-based techniques and furthered the work by [46]. In a similar setting, [29] studied the excess risk obtained by running gradient descent, and explored the double descent phenomenon with a peak around the separability threshold. The CGMT has been used in several other contexts, both in high-dimensional statistics and information theory, for example, to characterize the performance of the SLOPE estimator in sparse linear regression [52], to study high-dimensional regularized estimators in logistic regression [85], and to establish performance guarantees for PhaseMax [30]. The CGMT has proved useful in the study of high-dimensional convex problems, since it decouples a complex Gaussian process defined by a min-max objective function to a much simpler Gaussian process with essentially the same limit, yet much easier to analyze. However, this is merely a starting point or a basic building block. The study of the reduced optimization problem is problem-specific and is usually rather challenging in most high-dimensional settings, often requiring the development of nontrivial probabilistic analysis (see Appendix A for specific details in our case).

*Min-norm interpolation*. This paper investigates the min-$\ell_1$-norm interpolated classifier, which characterizes the limit of the Boosting solution on separable data. In recent years, min-norm interpolated solutions and their statistical properties have been extensively studied; see [5, 9–11, 23, 51, 63, 64, 67] for the regression problem, and [25, 29, 65, 75] for the classification problem. It has been conjectured that the implicit "min-norm" regularization, a

version of the Occam's razor principle, is responsible for the superior statistical behavior of complex over-parametrized models [10, 63, 106]. To the best of our knowledge, the current paper is the first to provide sharp statistical results for interpolated classifiers induced by the $\ell_1$ geometry (rather than the $\ell_2$), which has been argued to be a more suitable geometry [3, 4, 26, 33, 50] for the limit of gradient flow on shallow neural networks with 2-homogenous activations. In this light, we expect our results to be of much broader utility beyond the context of boosting.

**5. Discussion.** This paper establishes a high-dimensional asymptotic theory for AdaBoost and develops precise characterizations for both its generalization and optimization properties. This is achieved through an in-depth study of the max-$\ell_1$-margin, the min-$\ell_1$-norm interpolant and a sharp analysis of the time necessary for AdaBoost to approximate this interpolant arbitrarily well. In doing so, this work identifies the exact quantities that govern the generalization behavior of AdaBoost for a class of data-generation models, and the relationship between this test error and the optimal Bayes error. On the optimization front, we further uncover how overparametrization leads to faster optimization. The proposed theory demonstrates commendable finite sample behavior, applies for a broad class of statistical models, and is empirically robust to violations of certain assumptions. Natural variants of AdaBoost that correspond to max-$\ell_q$-margins for $q > 1$, are further analyzed.

We conclude with a couple of directions of future research: it would be of interest (a) to rigorously characterize analogous properties of AdaBoost for covariate distributions with arbitrary correlations; this is a particularly challenging task for general $\ell_q$ geometry when $q \neq 2$, as explained in Appendix B, and (b) to complement such characterizations via data-driven schemes for estimating the parameters $c_1^\star$, $c_2^\star$ that govern properties of the $\ell_1$ margin and interpolant, as well as the generalization performance of AdaBoost. Such estimation schemes are expected to be useful for providing recommendations regarding algorithm choice to practitioners.

**Acknowledgments.** T. Liang wishes to thank Yoav Freund, Bin Yu, Misha Belkin, as well as participants in the Learning Theory seminar at Google Research and NSF-Simons Collaboration on Mathematics of Deep Learning for constructive feedback that greatly improved the paper.

P. Sur wishes to thank the organizers and participants of the Young Data Science Researcher Seminar, ETH Zurich, for constructive feedback.

Both authors gratefully thank the anonymous referees and the associate editor who provided helpful comments that greatly improved the manuscript.

**Funding.** T. Liang acknowledges support from the NSF Career award (DMS-2042473), the George C. Tiao faculty fellowship and the William S. Fishman faculty research fund at the University of Chicago Booth School of Business.

P. Sur was partially supported by the Center for Research on Computation and Society, Harvard John A. Paulson School of Engineering and Applied Sciences and by NSF DMS-2113426.

## SUPPLEMENTARY MATERIAL

**Supplement to "A precise high-dimensional asymptotic theory for boosting and minimum-$\ell_1$-norm interpolated classifiers"** (DOI: 10.1214/22-AOS2170SUPP; .pdf). The supplementary material contains all the proofs and a discussion regarding possibilities of extending our results to feature covariances not covered in this manuscript. This can be found at doi: TBA.

## REFERENCES


[1] ADVANI, M. and GANGULI, S. (2016). Statistical mechanics of optimal convex inference in high dimensions. *Phys. Rev. A* **6** 031034.

[2] ALBERT, A. and ANDERSON, J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika* **71** 1–10. MR0738319 https://doi.org/10.1093/biomet/71.1.1

[3] AMID, E. and WARMUTH, M. K. (2020). Winnowing with gradient descent. *Proc. Mach. Learn. Res.* **125** 1–20.

[4] BACH, F. (2017). Breaking the curse of dimensionality with convex neutral networks. *J. Mach. Learn. Res.* **18** 19. MR3634886

[5] BARTLETT, P. L., LONG, P. M., LUGOSI, G. and TSIGLER, A. (2020). Benign overfitting in linear regression. *Proc. Natl. Acad. Sci. USA* **117** 30063–30070. MR4263288 https://doi.org/10.1073/pnas.1907378117

[6] BARTLETT, P. L. and TRASKIN, M. (2007). AdaBoost is consistent. *J. Mach. Learn. Res.* **8** 2347–2368. MR2353835

[7] BEAN, D., BICKEL, P. J., EL KAROUI, N. and YU, B. (2013). Optimal M-estimation in high-dimensional regression. *Proc. Natl. Acad. Sci. USA* **110** 14563–14568.

[8] BELKIN, M., HSU, D., MA, S. and MANDAL, S. (2018). Reconciling modern machine learning and the bias-variance trade-off. ArXiv preprint. Available at arXiv:1812.11118.

[9] BELKIN, M., HSU, D. and XU, J. (2020). Two models of double descent for weak features. *SIAM J. Math. Data Sci.* **2** 1167–1180. MR4186534 https://doi.org/10.1137/20M1336072

[10] BELKIN, M., MA, S. and MANDAL, S. (2018). To understand deep learning we need to understand kernel learning. ArXiv preprint. Available at arXiv:1802.01396.

[11] BELKIN, M., RAKHLIN, A. and TSYBAKOV, A. B. (2018). Does data interpolation contradict statistical optimality? ArXiv preprint. Available at arXiv:1806.09471.

[12] BICKEL, P. J., RITOV, Y. and ZAKAI, A. (2006). Some theory for generalized Boosting Algorithms. *J. Mach. Learn. Res.* **7** 705–732. MR2274384

[13] BLANCHARD, G., LUGOSI, G. and VAYATIS, N. (2004). On the rate of convergence of regularized boosting classifiers. *J. Mach. Learn. Res.* **4** 861–894. MR2076000 https://doi.org/10.1162/1532443041424319

[14] BREIMAN, L. (1996). Arcing classifiers. *Ann. Statist.* **26** 123–40.

[15] BREIMAN, L. (1996). Bias, variance, and arcing classifiers Technical Report Tech. Rep. 460, Statistics Department, Univ. California, Berkeley.

[16] BREIMAN, L. (1999). Prediction games and arcing algorithms. *Neural Comput.* **11** 1493–1517.

[17] BREIMAN, L. (2004). Population theory for boosting ensembles. *Ann. Statist.* **32** 1–11. MR2050998 https://doi.org/10.1214/aos/1079120126

[18] BÜHLMANN, P. (2006). Boosting for high-dimensional linear models. *Ann. Statist.* **34** 559–583. MR2281878 https://doi.org/10.1214/009053606000000092

[19] BÜHLMANN, P. and HOTHORN, T. (2007). Boosting Algorithms: Regularization, prediction and model fitting. *Statist. Sci.* **22** 477–505. MR2420454 https://doi.org/10.1214/07-STS242

[20] BÜHLMANN, P. and HOTHORN, T. (2010). Twin boosting: Improved feature selection and prediction. *Stat. Comput.* **20** 119–138. MR2610767 https://doi.org/10.1007/s11222-009-9148-5

[21] BÜHLMANN, P. and YU, B. (2003). Boosting with the $L_2$ loss: Regression and classification. *J. Amer. Statist. Assoc.* **98** 324–339. MR1995709 https://doi.org/10.1198/016214503000125

[22] BÜHLMANN, P. and YU, B. (2006). Sparse boosting. *J. Mach. Learn. Res.* **7** 1001–1024. MR2274395

[23] BUNEA, F., STRIMAS-MACKEY, S. and WEGKAMP, M. (2020). Interpolation under latent factor regression models. ArXiv preprint. Available at arXiv:2002.02525.

[24] CANDÈS, E. J. and SUR, P. (2020). The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression. *Ann. Statist.* **48** 27–42. MR4065151 https://doi.org/10.1214/18-AOS1789

[25] CHATTERJI, N. S. and LONG, P. M. (2021). Finite-sample analysis of interpolating linear classifiers in the overparameterized regime. *J. Mach. Learn. Res.* **22** 129. MR4279780

[26] CHIZAT, L. and BACH, F. (2020). Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. In *Conference on Learning Theory* 1305–1338. PMLR.

[27] CHOULDECHOVA, A. and ROTH, A. (2018). The frontiers of fairness in machine learning. ArXiv preprint. Available at arXiv:1810.08810.

[28] COVER, T. M. (1965). Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Trans. Electron. Comput.* **3** 326–334.

[29] DENG, Z., KAMMOUN, A. and THRAMPOULIDIS, C. (2019). A model of double descent for high-dimensional binary linear classification. Available at arXiv:1911.05822 [cs, Eess, Stat].

[30] DHIFALLAH, O., THRAMPOULIDIS, C. and LU, Y. M. (2018). Phase retrieval via polytope optimization: Geometry, phase transitions, and new algorithms. ArXiv preprint. Available at arXiv:1805.09555.
[31] DONOHO, D. and MONTANARI, A. (2016). High dimensional robust M-estimation: Asymptotic variance via approximate message passing. *Probab. Theory Related Fields* **166** 935–969. MR3568043 https://doi.org/10.1007/s00440-015-0675-z
[32] DONOHO, D. L., MALEKI, A. and MONTANARI, A. (2009). Message-passing algorithms for compressed sensing. *Proc. Natl. Acad. Sci. USA* **106** 18914–18919.
[33] DOU, X. and LIANG, T. (2021). Training neural networks as learning data-adaptive kernels: Provable representation and approximation benefits. *J. Amer. Statist. Assoc.* **116** 1507–1520. MR4309289 https://doi.org/10.1080/01621459.2020.1745812
[34] DRUCKER, H. and CORTES, C. (1996). Boosting decision trees. In *Advances in Neural Information Processing Systems* 479–485.
[35] EL KAROUI, N. (2018). On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators. *Probab. Theory Related Fields* **170** 95–175. MR3748322 https://doi.org/10.1007/s00440-016-0754-9
[36] EL KAROUI, N., BEAN, D., BICKEL, P. J., LIM, C. and YU, B. (2013). On robust regression with high-dimensional predictors. *Proc. Natl. Acad. Sci. USA* **110** 14557–14562.
[37] FENG, O. Y., VENKATARAMANAN, R., RUSH, C. and SAMWORTH, R. J. (2021). A unifying tutorial on approximate message passing. ArXiv preprint. Available at arXiv:2105.02180.
[38] FREUND, R. M., GRIGAS, P. and MAZUMDER, R. (2013). Adaboost and forward stagewise regression are first-order convex optimization methods. ArXiv preprint. Available at arXiv:1307.1192.
[39] FREUND, R. M., GRIGAS, P. and MAZUMDER, R. (2017). A new perspective on boosting in linear regression via subgradient optimization and relatives. *Ann. Statist.* **45** 2328–2364. MR3737894 https://doi.org/10.1214/16-AOS1505
[40] FREUND, Y. (1995). Boosting a weak learning algorithm by majority. *Inform. and Comput.* **121** 256–285. MR1348530 https://doi.org/10.1006/inco.1995.1136
[41] FREUND, Y. and SCHAPIRE, R. E. (1995). A decision-theoretic generalization of on-line learning and an application to boosting. In *European Conference on Computational Learning Theory* 23–37. Springer, Berlin.
[42] FREUND, Y. and SCHAPIRE, R. E. (1995). A desicion-theoretic generalization of on-line learning and an application to boosting. In *European Conference on Computational Learning Theory* 23–37. Springer, Berlin.
[43] FREUND, Y. and SCHAPIRE, R. E. (1996). Experiments with a new boosting algorithm. In *Icml* **96** 148–156. Citeseer.
[44] FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2000). Additive logistic regression: A statistical view of boosting (with discussion and a rejoinder by the authors). *Ann. Statist.* **28** 337–407. MR1790002 https://doi.org/10.1214/aos/1016218223
[45] FRIEDMAN, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Ann. Statist.* **29** 1189–1232. MR1873328 https://doi.org/10.1214/aos/1013203451
[46] GARDNER, E. (1988). The space of interactions in neural network models. *J. Phys. A* **21** 257–270. MR0939730
[47] GORDON, Y. (1985). Some inequalities for Gaussian processes and applications. *Israel J. Math.* **50** 265–289. MR0800188 https://doi.org/10.1007/BF02759761
[48] GORDON, Y. (1988). On Milman's inequality and random subspaces which escape through a mesh in $\mathbf{R}^n$. In *Geometric Aspects of Functional Analysis* (1986/87). *Lecture Notes in Math.* **1317** 84–106. Springer, Berlin. MR0950977 https://doi.org/10.1007/BFb0081737
[49] GROVE, A. J. and SCHUURMANS, D. (1998). Boosting in the limit: Maximizing the margin of learned ensembles. In *AAAI/IAAI* 692–699.
[50] GUNASEKAR, S., LEE, J., SOUDRY, D. and SREBRO, N. (2018). Characterizing implicit bias in terms of optimization geometry. In *International Conference on Machine Learning* 1832–1841. PMLR.
[51] HASTIE, T., MONTANARI, A., ROSSET, S. and TIBSHIRANI, R. J. (2019). Surprises in high-dimensional ridgeless least squares interpolation. ArXiv preprint. Available at arXiv:1903.08560.
[52] HU, H. and LU, Y. M. (2019). Asymptotics and optimal designs of SLOPE for sparse linear regression. In 2019 *IEEE International Symposium on Information Theory* (*ISIT*) 375–379. IEEE, Los Alamitos.
[53] HU, H. and LU, Y. M. (2020). Universality laws for high-dimensional learning with random features. ArXiv preprint. Available at arXiv:2009.07669.
[54] JI, Z. and TELGARSKY, M. (2021). Characterizing the implicit bias via a primal-dual analysis. In *Algorithmic Learning Theory* 772–804. PMLR.
[55] JIANG, W. (2001). Some theoretical aspects of boosting in the presence of noisy data. In *Proceedings of the Eighteenth International Conference on Machine Learning* Citeseer.

[56] JIANG, W. (2004). Process consistency for AdaBoost. *Ann. Statist.* **32** 13–29. MR2050999 https://doi.org/10.1214/aos/1079120128

[57] KLEINBERG, J. and MULLAINATHAN, S. (2019). Simplicity creates inequity: Implications for fairness, stereotypes, and interpretability. In *Proceedings of the* 2019 *ACM Conference on Economics and Computation* 807–808.

[58] KOLTCHINSKII, V. and BEZNOSOVA, O. (2005). Exponential convergence rates in classification. In *Learning Theory. Lecture Notes in Computer Science* **3559** 295–307. Springer, Berlin. MR2203269 https://doi.org/10.1007/11503415_20

[59] KOLTCHINSKII, V. and PANCHENKO, D. (2002). Empirical margin distributions and bounding the generalization error of combined classifiers. *Ann. Statist.* **30** 1–50. MR1892654 https://doi.org/10.1214/aos/1015362182

[60] KOLTCHINSKII, V. and PANCHENKO, D. (2005). Complexities of convex combinations and bounding the generalization error in classification. *Ann. Statist.* **33** 1455–1496. MR2166553 https://doi.org/10.1214/009053605000000228

[61] LESAFFRE, E. and ALBERT, A. (1989). Partial separation in logistic discrimination. *J. Roy. Statist. Soc. Ser. B* **51** 109–116. MR0984997

[62] LI, A. H. and BRADIC, J. (2018). Boosting in the presence of outliers: Adaptive classification with nonconvex loss functions. *J. Amer. Statist. Assoc.* **113** 660–674. MR3832217 https://doi.org/10.1080/01621459.2016.1273116

[63] LIANG, T. and RAKHLIN, A. (2020). Just interpolate: Kernel "ridgeless" regression can generalize. *Ann. Statist.* **48** 1329–1347. MR4124325 https://doi.org/10.1214/19-AOS1849

[64] LIANG, T., RAKHLIN, A. and ZHAI, X. (2020). On the multiple descent of minimum-norm interpolants and restricted lower isometry of kernels. In *Proceedings of* 33*rd Conference on Learning Theory* (J. Abernethy and S. Agarwal, eds.). *Proceedings of Machine Learning Research* **125** 2683–2711. PMLR.

[65] LIANG, T. and RECHT, B. (2021). Interpolating classifiers make few mistakes. ArXiv preprint. Available at arXiv:2101.11815.

[66] LIANG, T. and SUR, P. (2022). Supplement to "A Precise High-Dimensional Asymptotic Theory for Boosting and Minimum-$\ell_1$-Norm Interpolated Classifiers." https://doi.org/10.1214/22-AOS2170SUPP

[67] LIANG, T. and TRAN-BACH, H. (2021). Mehler's formula, branching process, and compositional kernels of deep neural networks. *J. Amer. Statist. Assoc.* **0** 1–14.

[68] LIPTON, Z. C. (2018). The mythos of model interpretability. *ACM Queue* **16** 31–57.

[69] LUGOSI, G. and VAYATIS, N. (2004). On the Bayes-risk consistency of regularized boosting methods. *Ann. Statist.* **32** 30–55. MR2051000 https://doi.org/10.1214/aos/1079120129

[70] MANNOR, S. and MEIR, R. (2001). Geometric bounds for generalization in boosting. In *Computational Learning Theory* (*Amsterdam*, 2001). *Lecture Notes in Computer Science* **2111** 461–472. Springer, Berlin. MR2042053 https://doi.org/10.1007/3-540-44581-1_30

[71] MANNOR, S. and MEIR, R. (2002). On the existence of linear weak learners and applications to boosting. *Mach. Learn.* **48** 219–251.

[72] MANNOR, S., MEIR, R. and ZHANG, T. (2002). The consistency of greedy algorithms for classification. In *Computational Learning Theory* (*Sydney*, 2002). *Lecture Notes in Computer Science* **2375** 319–333. Springer, Berlin. MR2040422 https://doi.org/10.1007/3-540-45435-7_22

[73] MASON, L., BAXTER, J., BARTLETT, P. L. and FREAN, M. R. (2000). Boosting algorithms as gradient descent. In *Advances in Neural Information Processing Systems* 512–518.

[74] MEI, S. and MONTANARI, A. (2019). The generalization error of random features regression: Precise asymptotics and the double descent curve. *Comm. Pure Appl. Math.*

[75] MONTANARI, A., RUAN, F., SOHN, Y. and YAN, J. (2019). The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime. ArXiv preprint. Available at arXiv:1911.01544.

[76] MUKHERJEE, I., RUDIN, C. and SCHAPIRE, R. E. (2011). The rate of convergence of AdaBoost. In *Proceedings of the* 24*th Annual Conference on Learning Theory* 537–558.

[77] QUINLAN, J. (1996). Bagging, boosting, and C4. 5. In 'AAAI'96 Proceedings of the Thirteenth National Conference on Artificial Intelligence–Volume 1', 4–8 August 1996, Portland, OR, USA.

[78] RAHIMI, A. and RECHT, B. (2009). Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. In *Advances in Neural Information Processing Systems* 21 (D. Koller, D. Schuurmans, Y. Bengio and L. Bottou, eds.) 1313–1320. Curran Associates, Red Hook.

[79] RÄTSCH, G., ONODA, T. and MÜLLER, K.-R. (2001). Soft margins for AdaBoost. *Mach. Learn.* **42** 287–320.

[80] RÄTSCH, G. and WARMUTH, M. K. (2005). Efficient margin maximizing with boosting. *J. Mach. Learn. Res.* **6** 2131–2152. MR2249883

[81] REYZIN, L. and SCHAPIRE, R. E. (2006). How boosting the margin can also boost classifier complexity. In *Proceedings of the* 23*rd International Conference on Machine Learning* 753–760.
[82] ROSSET, S., ZHU, J. and HASTIE, T. (2003/04). Boosting as a regularized path to a maximum margin classifier. *J. Mach. Learn. Res.* **5** 941–973. MR2248005
[83] RUDIN, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **1** 206–215.
[84] RUDIN, C., SCHAPIRE, R. E. and DAUBECHIES, I. (2007). Analysis of Boosting Algorithms using the smooth margin function. *Ann. Statist.* **35** 2723–2768. MR2382664 https://doi.org/10.1214/009053607000000785
[85] SALEHI, F., ABBASI, E. and HASSIBI, B. (2019). The impact of regularization on high-dimensional logistic regression. In *Advances in Neural Information Processing Systems* 11982–11992.
[86] SANTNER, T. J. and DUFFY, D. E. (1986). A note on A. Albert and J. A. Anderson's conditions for the existence of maximum likelihood estimates in logistic regression models. *Biometrika* **73** 755–758. MR0897873 https://doi.org/10.1093/biomet/73.3.755
[87] SCHAPIRE, R. E. (1990). The strength of weak learnability. *Mach. Learn.* **5** 197–227.
[88] SCHAPIRE, R. E., FREUND, Y., BARTLETT, P. and LEE, W. S. (1998). Boosting the margin: A new explanation for the effectiveness of voting methods. *Ann. Statist.* **26** 1651–1686. MR1673273 https://doi.org/10.1214/aos/1024691352
[89] SHALEV-SHWARTZ, S. and SINGER, Y. (2010). On the equivalence of weak learnability and linear separability: New relaxations and efficient Boosting Algorithms. *Mach. Learn.* **80** 141–163. MR3108163 https://doi.org/10.1007/s10994-010-5173-z
[90] SHCHERBINA, M. and TIROZZI, B. (2003). Rigorous solution of the Gardner problem. *Comm. Math. Phys.* **234** 383–422. MR1964377 https://doi.org/10.1007/s00220-002-0783-3
[91] STOJNIC, M. (2013). A framework to characterize performance of lasso algorithms. ArXiv preprint. Available at arXiv:1303.7291.
[92] STOJNIC, M. (2013). Meshes that trap random subspaces. ArXiv preprint. Available at arXiv:1304.0003.
[93] STOJNIC, M. (2013). Upper-bounding l1-optimization weak thresholds. Available at ArXiv.
[94] SUR, P. (2019). *A Modern Maximum Likelihood Theory for High-Dimensional Logistic Regression*. ProQuest LLC, Ann Arbor, MI. Thesis (Ph.D.)—Stanford University. MR4197622
[95] SUR, P. and CANDÈS, E. J. (2019). A modern maximum-likelihood theory for high-dimensional logistic regression. *Proc. Natl. Acad. Sci. USA* **116** 14516–14525. MR3984492 https://doi.org/10.1073/pnas.1810420116
[96] SUR, P., CHEN, Y. and CANDÈS, E. J. (2019). The likelihood ratio test in high-dimensional logistic regression is asymptotically a *rescaled* chi-square. *Probab. Theory Related Fields* **175** 487–558. MR4009715 https://doi.org/10.1007/s00440-018-00896-9
[97] TAHERI, H., PEDARSANI, R. and THRAMPOULIDIS, C. (2020). Sharp asymptotics and optimal performance for inference in binary models. In *International Conference on Artificial Intelligence and Statistics* 3739–3749. PMLR.
[98] TELGARSKY, M. (2013). Margins, shrinkage, and boosting. ArXiv preprint. Available at arXiv:1303.4172.
[99] THRAMPOULIDIS, C., ABBASI, E. and HASSIBI, B. (2015). Lasso with non-linear measurements is equivalent to one with linear measurements. *Adv. Neural Inf. Process. Syst.* **28** 3420–3428.
[100] THRAMPOULIDIS, C., ABBASI, E. and HASSIBI, B. (2018). Precise error analysis of regularized $M$-estimators in high dimensions. *IEEE Trans. Inf. Theory* **64** 5592–5628. MR3832326 https://doi.org/10.1109/TIT.2018.2840720
[101] THRAMPOULIDIS, C., OYMAK, S. and HASSIBI, B. (2014). The Gaussian min-max theorem in the presence of convexity. ArXiv preprint. Available at arXiv:1408.4837.
[102] THRAMPOULIDIS, C., OYMAK, S. and HASSIBI, B. (2015). Regularized linear regression: A precise analysis of the estimation error. In *Conference on Learning Theory* 1683–1709.
[103] VILLANI, C. (2008). *Optimal Transport*: *Old and New* **338**. Springer, Berlin.
[104] WANG, S., WENG, H. and MALEKI, A. (2020). Which bridge estimator is the best for variable selection? *Ann. Statist.* **48** 2791–2823. MR4152121 https://doi.org/10.1214/19-AOS1906
[105] WELLER, A. (2019). Transparency: Motivations and challenges. In *Explainable AI*: *Interpreting*, *Explaining and Visualizing Deep Learning* 23–40. Springer, Berlin.
[106] ZHANG, C., BENGIO, S., HARDT, M., RECHT, B. and VINYALS, O. (2016). Understanding deep learning requires rethinking generalization. ArXiv preprint. Available at arXiv:1611.03530.
[107] ZHANG, T. (2004). Statistical behavior and consistency of classification methods based on convex risk minimization. *Ann. Statist.* **32** 56–85. MR2051001 https://doi.org/10.1214/aos/1079120130
[108] ZHANG, T. and YU, B. (2005). Boosting with early stopping: Convergence and consistency. *Ann. Statist.* **33** 1538–1579. MR2166555 https://doi.org/10.1214/009053605000000255
[109] ZHAO, Q., SUR, P. and CANDES, E. J. (2020). The asymptotic distribution of the MLE in high-dimensional logistic models: Arbitrary covariance. ArXiv preprint. Available at arXiv:2001.09351.