WORKING PAPER · NO. 2020-152

A Precise High-Dimensional Asymptotic Theory for Boosting and Minimum-ℓ_1-Norm Interpolated Classifiers

Tengyuan Liang and Pragya Sur

OCTOBER 2020

Becker Friedman Institute | Big Data Initiative

5757 S. University Ave.
Chicago, IL 60637
Main: 773.702.5599
bfi.uchicago.edu

A Precise High-Dimensional Asymptotic Theory for Boosting and Minimum-ℓ_1-Norm Interpolated Classifiers

Tengyuan Liang [*1] and Pragya Sur [†2]

[1]University of Chicago
[2]Harvard University

Abstract

This paper establishes a precise high-dimensional asymptotic theory for boosting on separable data, taking statistical and computational perspectives. We consider the setting where the number of features (weak learners) p scales with the sample size n, in an over-parametrized regime. Under a broad 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 ℓ_1-margin. The relation between the boosting test error and the optimal Bayes error is pinned down explicitly. In turn, these precise characterizations resolve several open questions raised in [15, 81] surrounding boosting. On the computational front, we provide a sharp analysis of the stopping time when boosting approximately maximizes the empirical ℓ_1 margin. Furthermore, we discover that the larger the overparametrization ratio p/n, the smaller the proportion of active features (with zero initialization), and the faster the optimization reaches interpolation. At the heart of our theory lies an in-depth study of the maximum ℓ_1-margin, which can be accurately described by a new system of non-linear equations; we analyze this margin and the properties of this system, using Gaussian comparison techniques and a novel uniform deviation argument. Variants of AdaBoost corresponding to general ℓ_q geometry, for $q > 1$, are also presented, together with an exact analysis of the high-dimensional generalization and optimization behavior of a class of these algorithms.

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 over-parametrized 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

[*]tengyuan.liang@chicagobooth.edu.
[†]pragya@fas.harvard.edu.

norms of the classifier. This paper focuses on the celebrated AdaBoost/boosting algorithm in this minimum-norm interpolation regime, where we conduct a precise analysis of its statistical and computational properties.

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 combine weak learners using intelligent schemes and exhibit remarkable generalization performance. The groundbreaking AdaBoost paper, Freund and Schapire [38], is widely regarded as the milestone in the boosting literature, which can be traced back even earlier [80, 37]. 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 [81], which led to decades of research activities devoted to a better understanding of this algorithm and its variants.

The seminal papers [14, 33, 71] 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 [95], and motivated the study of interpolation and implicit regularization [8, 7, 57, 48, 6, 59]. This peculiar and seemingly counter-intuitive 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 [81] 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. [81] 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, one being the fraction of training examples with margin below κ, and the other, $\kappa^{-1}\mathrm{C}(\mathcal{H})/\sqrt{n}$, involving the complexity of the class $\mathrm{C}(\mathcal{H})$ and the sample size n scaled by κ. 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, [81] commented that the proposed upper bounds can be sub-optimal 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 [15] 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 [53] improved the earlier bounds from [81]. Since all of these results involved upper bounds, a proper understanding of AdaBoost was still far from achieved; in fact Breiman in his Wald lectures (2002) commented that he would characterize *"the understanding of Adaboost as the most important open problem in machine learning."*

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 [16], who studied convergence properties of the algorithm in the population case. The seminal papers Jiang [51], Lugosi and Vayatis [61], Zhang [96] considered different

function classes and variants of boosting, and furthered this direction of research. [51] 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 [41, 46, 73, 65]—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 [97], Bartlett and Traskin [5]. 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-ℓ_1-norm interpolation (and implications). In a venture to understand the path of boosting iterates better, Rosset et al. [75], Zhang and Yu [97] established that for linearly separable data, AdaBoost with infinitesimal step size converges to the minimum ℓ_1-norm interpolated classifier (Equation (1.2)) when left to run forever. This interpolant is crucially related to the maximum ℓ_1-margin on the data, κ_{n,ℓ_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 [15], this leads to a plausible conjecture that the max-ℓ_1-margin is a crucial quantity that determines both generalization and optimization behaviors of boosting algorithms. (See also [89], 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-ℓ_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 missing from the existing literature.

Rosset, Zhu, and Hastie [75] further discussed that the aforementioned convergence to min-ℓ_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 [60, 26, 76, 52, 94]. Boosting has subsequently witnessed widespread development, and varying perspectives have emerged through several seminal works e.g. [41, 42, 20, 17, 77, 36]; see Section 4 for further discussions.

This paper. Prior literature suggested that the min-ℓ_1-norm interpolated classifier and the max-ℓ_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 with growing n.

Furthermore, boosting has empirically demonstrated exceptional performance in high dimensions, yet a precise understanding of AdaBoost in overparametrized settings has so far been out of reach. Therefore, to study properties of AdaBoost on separable data, it is both theoretically

necessary and empirically natural to analyze the algorithm in a high-dimensional setting. This paper addresses these crucial questions surrounding AdaBoost, in high dimensions, focussing on the case of binary classifications. 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 AdaBoost 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 [23, 67], as argued above; this regime has also been investigated for regression problems and other contexts (see for instance, [66, 35, 34, 31, 87, 88], and the references cited therein) and is well-known to produce asymptotic predictions with remarkable finite sample performance. Since we are primarily interested in overparametrized settings, we assume that the data is (asymptotically) linearly separable in the sense of Eqn. (2.6). (This is equivalent to the dimensionality ψ lying above a threshold that depends on the underlying signal strength of the problem [23, 67]; see Section 2 for further details.) Define the *min-ℓ_1-norm interpolated classifier* to be

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

Note that the min-ℓ_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-ℓ_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 κ_{n,ℓ_1} is positive. Then, under a host of different models (Sections 2 and 3.5), this paper provides the following contributions to the statistical and computational understanding of boosting:

- We characterize precisely the value of the max-ℓ_1-margin (Theorem 3.1) in the high-dimensional regime (1.1), answering a fundamental question raised in Breiman [15]. Informally, we show that $\sqrt{p}\kappa_{n,\ell_1}$ converges almost surely to a constant $\kappa_\star$ that depends on ψ 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 non-linear system of equations that we will introduce in (3.11). This limiting characterization will prove crucial for understanding the properties of AdaBoost on (asymptotically) separable data.

- In parallel, we establish precise formulae for the generalization error of the min-ℓ_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 ψ and the limit $\kappa_\star$ characterized in the preceding step. The consequences of this result for boosting will be discussed soon; notably, the min-ℓ_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.

- Turning to boosting, we provide a sharp characterization of a threshold T such that for all iterations $t \geq T$, the AdaBoost 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 3.4). Together with Theorems 3.1-3.2, this result immediately provides an exact characterization of the generalization error of boosting, and improves upon the existing upper bounds [81, 53] by a margin. This formula crucially involves $\kappa_\star$, and therefore, our results resolve an open question posed in Schapire et al. [81], Breiman [15] regarding which quantity truly governs the generalization performance of AdaBoost. Furthermore, the formula encodes a concrete recipe for comparing AdaBoost's test error with the Bayes error in high dimensions.

- The iteration threshold T from the previous step can be described through a precise formula (in the large n,p limit) that involves the limit of the max-ℓ_1-margin $\kappa_\star$. Utilizing this, we demonstrate two curious phenomena regarding overparametrization, both not known earlier for AdaBoost. (1) Keeping other problem parameters fixed, T decreases with an increase in ψ, 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. We demonstrate that this fraction can be small, formalizing the intuition that boosting should converge to a sparse classifier [75], and decreases as ψ increases, suggesting that *overparametrization leads to a classifier with a simpler interpretation*.

- Finally, we introduce a new class of boosting algorithms that converge to the max-ℓ_q-margin direction (Section 3.4) for $q > 1$. Rosset et al. [75] discussed the importance of studying such notions of margins, since it is unclear which geometry induces a better solution. Here, we construct such algorithms and provide precise analysess of their generalization (for the case $1 < q \leq 2$) and optimization properties (for all $q > 1$) in a spirit similar to that for AdaBoost done above.

On the theoretical end, our analysis builds upon classical results in Gaussian comparison inequalities [44, 45] that have been strengthened relatively recently [83, 90, 92], leading to the *Convex Gaussian Min-Max Theorem* (CGMT) (see Section 4 for a discussion). The topic of max-ℓ_2-margin has received considerable attention, dating back to [43, 82], and has more recently been analyzed in [67, 29]. Our proofs begin from these existing theory surrounding the max-ℓ_2-margin, particularly [67], however, the ℓ_2 and ℓ_q ($q \neq 2$) 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 Section 5.

Finite sample performance. Our results are asymptotic in nature, and here we test their applicability and accuracy in finite samples 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-ℓ_1-margin by two methods: (i) the numerical solution κ_{n,ℓ_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 *non-linear system of equations* defined in (3.11). This involved computing integrals, which

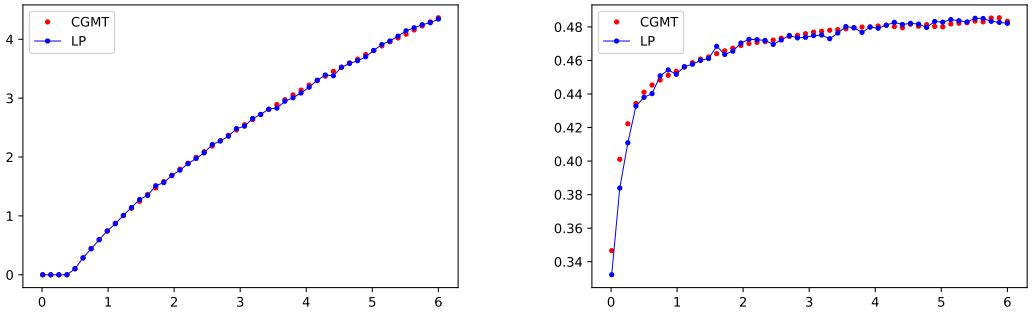

Figure 1: x-axis: Ratio p/n. y-axis: (a) Left: max-ℓ_1-margin, (b) Right: Generalization error.

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 $\mathbb{P}_{\mathbf{x},\mathbf{y}}\left(\mathbf{y}\cdot\mathbf{x}^\top\hat{\theta}_{n,\ell_1}<0\right)$, 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 cases, the points align remarkably well, demonstrating that our theory, albeit asymptotic, shows remarkable finite sample accuracy. In this example, the threshold for separability was around 0.43 [23]. This is also evidenced in the plot—the max-ℓ_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 5 (details are deferred to the Appendix). Section 4 discusses relevant literature that has been omitted from this introduction. Finally, Section 6 concludes with a discussion on possible directions for future work.

2 Formal setup and preliminaries

This section introduces our formal setup. Throughout, 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

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

Above, $\Lambda(n)\in\mathbb{R}^{p(n)\times p(n)}$ is a diagonal covariance matrix and f is any non-decreasing 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 delineated below. Conceptually, the structure of the problem is determined by four factors: overparametrization ψ, signal strength ρ, link function f, and a limiting measure μ 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 $c \le \lambda_i(n) \le 1/c$, $\forall 1 \le i \le p(n)$ 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*

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

where $e_{i,n}$ denotes the canonical vector in $\mathbb{R}^n$ 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 μ on $\mathbb{R}_{>0} \times \mathbb{R}$, in the Wasserstein-2 distance, that is,

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

which equivalently means weak convergence and convergence of the second moments (see for instance, [93, 67]). In particular, this implies that $\int w^2 \mu(d\lambda, dw) = 1$.

Assumption 3. *Finally, assume that*

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

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 [1], and this has been a problem of intense study in classical statistics and information theory [28, 79, 55]. More recently, [23] 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 [67]. 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 $\kappa \ge 0$,

$$F_\kappa(c_1, c_2) := \Big(\mathbb{E}\Big[(\kappa - c_1 Y Z_1 - c_2 Z_2)_+^2\Big]\Big)^{\frac{1}{2}} \quad \text{where} \begin{cases} Z_2 \perp (Y, Z_1) \\ Z_i \sim \mathcal{N}(0,1),\ i = 1, 2 \\ \mathbb{P}(Y = +1 | Z_1) = 1 - \mathbb{P}(Y = -1 | Z_1) = f(\rho \cdot Z_1) \end{cases} . \tag{2.7}$$

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 $F_\kappa : \mathbb{R} \times \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ 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 [40, 39]. Suppose that each weak learner outputs a binary decision $X_{ij} = x_i[j] \in \{-1, +1\}$ and $y_i \in \{-1, +1\}$. Let Δ_n be the standard probability simplex given by $\Delta_n := \{\boldsymbol{p} \in [0,1]^n : \sum_{i=1}^n p_i = 1\}$. AdaBoost at each step 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 $t \geq 0$:
   (a) Feature Selection: $v_{t+1} := \arg\min_{v \in \{\pm e_j\}_{j \in [p]}} \sum_{i \in [n]} \eta_t[i] \cdot \mathbb{I}_{y_i x_i^\top v \leq 0}$;
   (b) Adaptive Stepsize α_t: $\alpha_t = \frac{1}{2} \log\left(\frac{1 - \sum_{i \in [n]} \eta_t[i] \cdot \mathbb{I}_{y_i x_i^\top v_{t+1} \leq 0}}{\sum_{i \in [n]} \eta_t[i] \cdot \mathbb{I}_{y_i x_i^\top v_{t+1} \leq 0}} \right)$;
   (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 θ_T.

The *Boosting Algorithm* for continuous $X_{ij} = x_i[j] \in \mathbb{R}$ can be readily derived by modifying the steps 2.(a) and 2.(b) above. To be specific, choose the feature and the learning rate as follows

$$v_{t+1} := \arg\max_{v \in \{e_j\}_{j \in [p]}} |\eta_t^\top Z v|, \quad \alpha_t := \eta_t^\top Z v_{t+1}, \tag{2.9}$$

where $Z = y \circ X \in \mathbb{R}^{n \times p}$ denotes multiplying each element in the i-th row of X by y_i, $i \in [n]$.

3 Main Results

This section will provide precise analyses of the max-ℓ_1-margin κ_{n,ℓ_1} and the min-ℓ_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 (ψ, ρ, μ, f) introduced in Section 2.

3.1 Max-ℓ_1-margin and min-ℓ_1-norm interpolant

Recall the definition of the max-ℓ_1-margin from (1.3). We establish that κ_{n,ℓ_1}, when appropriately scaled, converges almost surely to a limit that can be explicitly characterized in terms of ψ, μ and f. To describe this limit, consider the following function first introduced in [67]: for any (ψ, κ) pair that satisfies $\psi > \psi^\downarrow(\kappa)$ (See Equation 3.13), define $T : (\psi, \kappa) \to \mathbb{R}$ to be

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

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 non-linear system of equations introduced in (3.11). A detailed description of this system is deferred until Section 3.2; the key point is that, the system takes as input the quantities ψ, ρ, μ, κ, and solves three equations in three unknowns, producing a triplet c_1, c_2, s. Throughout, μ and ρ 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 ψ, κ. It is helpful to remind the readers that we drop the dependence on f for simplicity of the exposition; however, it is essential 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-ℓ_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 above Theorem answers the fundamental question raised in Breiman [15] on the exact value of the max-min margin (1.3), for a wide range of problem instances. In fact, this max-ℓ_1-margin was conjectured to be a central quantity for boosting [15], and Theorem 3.1 provides a precise high-dimensional characterization of this object. It is remarkable to us that such almost sure convergence on this complex quantity crucial to boosting can be established. This limiting result will lead to extremely precise characterizations of statistical and computational properties of *Boosting Algorithms* in high dimensions, as we shall shortly see in Section 3.3. Although the result is asymptotic, the empirical margin (scaled) $\sqrt{p}\kappa_{n,\ell_1}$ shows remarkable agreement with the limiting value $\kappa_\star(\psi,\rho,\mu)$, even for datasets 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 towards this claim. Next, (3.3) clearly demonstrates the dependence of $\kappa_\star(\psi,\rho,\mu)$ on the overparametrization ratio ψ. Its dependence on the signal strength ρ and the distribution μ 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-ℓ_1-norm interpolated classifier (1.2), and its precise generalization behavior in our asymptotic regime (1.1). Define

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

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 (3.11), when the inputs to the system are ψ, ρ, μ 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 ρ.

Theorem 3.2. *Under the assumptions of Theorem 3.1, the generalization error of any min-ℓ_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,*

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

Theorem 3.2 provides an exact quantification of the generalization behavior of the min-ℓ_1-norm interpolant, which characterizes the long time and infinitesimal step size limit of AdaBoost on

separable data [75, 97]. Later, Section 3.3 will establish a further precise connection between $\hat{\theta}_{n,\ell_1}$ and the AdaBoost iterates (with appropriately chosen learning rates)—informally, the AdaBoost iterates arrive arbitrarily close to the min-ℓ_1-norm interpolant, beyond a certain time threshold. Therefore, Theorem 3.2 provides two important contributions to the boosting literature, described as follows.

First, an open question was posed by Schapire et al. [81], Breiman [15] 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-ℓ_1-margin precisely determines the generalization error. Since our result is asymptotically sharp (rather than merely an upper bound), Theorem 3.2 provides an answer to the question posed in [81, 15]. In contrast, existing margin-based generalization upper bounds [81, 53] scale as

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

which is loose compared to our sharp characterization. In fact, note that the inverse of the y-axis in Figure 3 corresponds to the classical upper bound $(\sqrt{n}\kappa_{n,\ell_1})^{-1}$ on the generalization error, as given by Eqn. (3.6), which is vacuous in this setting (even overlooking the log factors) since the upper bound is worse than 0.5. As a crucial remark, note that despite its asymptotic nature, Theorem 3.2 also exhibits remarkable finite sample performance, as already seen in Figure 1.

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

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

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, $\left\|\Pi_{(\Lambda^{1/2}\theta_\star)^\perp}(\Lambda^{1/2}\hat{\theta}_{n,\ell_1})\right\| \xrightarrow{\text{a.s.}} c_2^\star$. Recall the Bayes error formula, and contrast it with the test error formula (3.4) proved in Theorem 3.2,

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

Then, it is clear to see that $(c_2^\star)^{-1}c_1^\star$ exactly determines how the test error of $\hat{\theta}_{n,\ell_1}$ differs from the optimal Bayes error. Therefore, Theorem 3.2 significantly advances the literature on how the test error of boosting relates to the Bayes error [16, 51, 61, 96]: the optimality of Boosting (w.r.t. the optimal Bayes classifier) is entirely determined by the magnitude of $(c_2^\star)^{-1}c_1^\star$. In fact, one can estimate $(c_2^\star)^{-1}c_1^\star$ based on data to infer the optimality of boosting.

The constants $c_1^\star, c_2^\star, s^\star$ introduced in (3.4) and the discussion thereafter, form crucial components governing the behavior of $\hat{\theta}_{n,\ell_1}$. Theorem 3.2 demonstrated this through the lens of the generalization error. Additionally, we can characterize a multitude of other properties of $\hat{\theta}_{n,\ell_1}$.

Theorem 3.3. *Consider a convex bivariate function $\phi:\mathbb{R}^p\times\mathbb{R}^p\to\mathbb{R}$ that takes on a separable form $\phi(x,y)=\frac{1}{p}\sum_{j=1}^p\phi_0(\sqrt{p}x_j,\sqrt{p}y_j)$. Under the assumptions of Theorem 3.2, the min-ℓ_1-norm interpolant satisfies*

$$\phi(\hat{\theta}_{n,\ell_1}, \theta_\star) \xrightarrow{a.s.}$$
$$\mathbb{E}_{(\Lambda,W,G)}\phi_0\left(-\frac{\Lambda^{-1}\,\mathbf{prox}_{s^\star}\left(\Lambda^{1/2}G + \psi^{-1/2}[\partial_1 F_{\kappa_\star}(c_1^\star, c_2^\star) - c_1^\star c_2^{\star -1}\partial_2 F_{\kappa_\star}(c_1^\star, c_2^\star)]\Lambda^{1/2}W\right)}{\psi^{-1/2}c_2^{\star -1}\partial_2 F_{\kappa_\star}(c_1^\star, c_2^\star)}, \rho\Lambda^{-1/2}W\right), \tag{3.9}$$

where $(\Lambda, W, G) \sim \mu \otimes \mathcal{N}(0,1)$, *with* μ *defined as in* (2.4), $\kappa_\star = \kappa_\star(\psi, \rho, \mu)$ *from* (3.2), *and* $\mathbf{prox}_\lambda(\cdot)$ *is the proximal operator of the* ℓ_1 *norm, given by [70]*

$$\mathbf{prox}_\lambda(t) = \arg\min_s \left\{\lambda|s| + \frac{1}{2}(s-t)^2\right\} = \mathrm{sign}(t)\,(|t| - \lambda)_+ \quad . \tag{3.10}$$

Once again, in (3.9), $c_1^\star, c_2^\star, s^\star$ is the solution to (3.11) obtained on plugging in ψ, ρ, μ and $\kappa_\star(\psi, \rho, \mu)$. Theorem 3.3 can be used to characterize several general properties of the min-norm-interpolant, such as the distance of $\hat{\theta}_{n,\ell_1}$ from $\theta_\star$ in Euclidean norm (or in any p-norm), the empirical distribution of $\hat{\theta}_{n,\ell_1}$, and so on. Note that, the generalization error simply involved $c_1^\star, c_2^\star$, which could be interpreted as the angle and the orthogonal projection respectively. For general functions ϕ, $s^\star$ also features crucially in the limit (3.9), as the threshold for the soft-thresholding function (3.10). Now, $s^\star$ can be interpreted as the limit of certain Lagrange multipliers induced by the ℓ_1 constraint on the outer maximization in (1.3); this will be elaborated further in Section 5.1, Step 4.

The curious reader may wonder about the accuracy of our asymptotic theory for design matrices excluded from our assumptions. To test this sensitivity, we consider the setting of Figure 1 and calculate the max-ℓ_1-margin, based on the linear program (1.3), as well as the difference between the test error and Bayes error for two other classes of design matrices—(a) a Rademacher design where each entry of X is independently ± 1 with probability 1/2; (b) Gaussian designs where the covariance of each x_i is given by an AR(1) model, with auto-correlation $r = 0.1, 0.2, 0.3$. The left two subfigures compare the results for the Rademacher design versus the corresponding Gaussian design with matching second moments. Observe the match between the two settings, suggesting the applicability of our theory for a much broader class of covariate distributions. The right two subfigures display the AR(1) designs from Case (b), contrasted with $r = 0$. The variation across values of r is mild, and the shapes of the curves remain invariant, suggesting that our theory approximates the margin and test error values well for a class of weakly-correlated designs. We defer theoretical analysis beyond Gaussian designs to future work.

3.2 The non-linear system of equations

We will now introduce a non-linear system of equations that is key to the study of the max-ℓ_1-margin and the min-ℓ_1-norm interpolant in high dimensions, as delineated in Theorems 3.1–3.3.

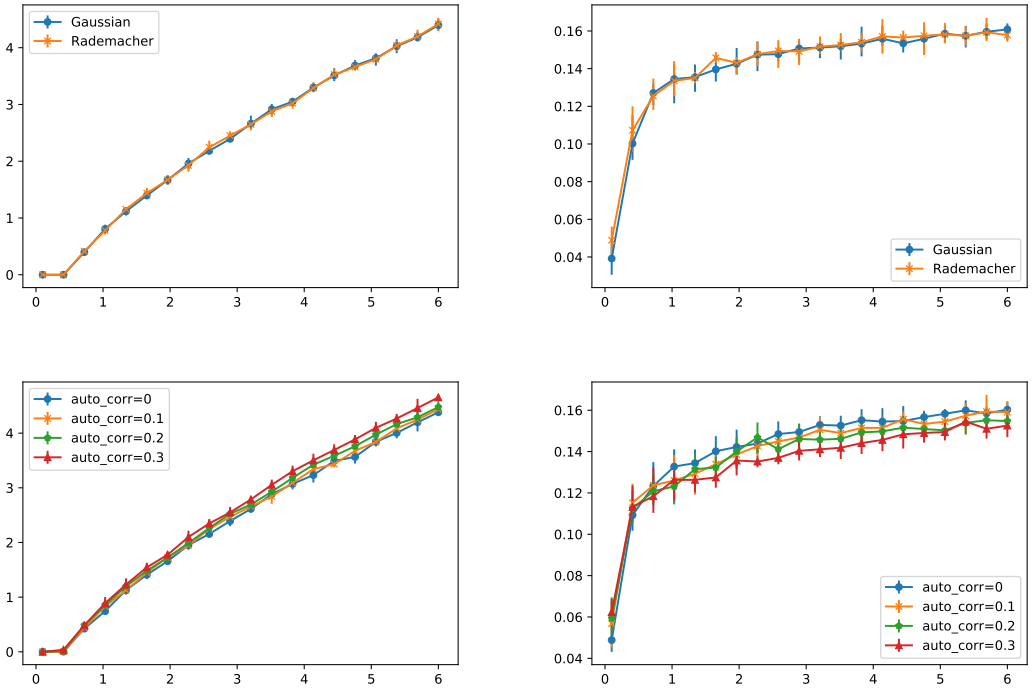

Figure 2: Subfigures (1) and (2) correspond to Gaussian design vs. Rademacher design, and subfigures (3) and (4) correspond to weakly-correlated design with auto-correlation $r = 0, 0.1, 0.2, 0.3$. x-axis: Ratio p/n. y-axis: (1) and (3) max-ℓ_1-margin, (2) and (4) Test error minus the Bayes error.

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

$$\begin{aligned} c_1 &= -\mathop{\mathbf{E}}_{(\Lambda,W,G)\sim\mathcal{Q}}\left(\frac{\Lambda^{-1/2}W\cdot\mathbf{prox}_s\left(\Lambda^{1/2}G+\psi^{-1/2}[\partial_1 F_\kappa(c_1,c_2)-c_1c_2^{-1}\partial_2 F_\kappa(c_1,c_2)]\Lambda^{1/2}W\right)}{\psi^{-1/2}c_2^{-1}\partial_2 F_\kappa(c_1,c_2)}\right) \\ c_1^2+c_2^2 &= \mathop{\mathbf{E}}_{(\Lambda,W,G)\sim\mathcal{Q}}\left(\frac{\Lambda^{-1/2}\,\mathbf{prox}_s\left(\Lambda^{1/2}G+\psi^{-1/2}[\partial_1 F_\kappa(c_1,c_2)-c_1c_2^{-1}\partial_2 F_\kappa(c_1,c_2)]\Lambda^{1/2}W\right)}{\psi^{-1/2}c_2^{-1}\partial_2 F_\kappa(c_1,c_2)}\right)^2 \quad . \quad (3.11) \\ 1 &= \mathop{\mathbf{E}}_{(\Lambda,W,G)\sim\mathcal{Q}}\left|\frac{\Lambda^{-1}\,\mathbf{prox}_s\left(\Lambda^{1/2}G+\psi^{-1/2}[\partial_1 F_\kappa(c_1,c_2)-c_1c_2^{-1}\partial_2 F_\kappa(c_1,c_2)]\Lambda^{1/2}W\right)}{\psi^{-1/2}c_2^{-1}\partial_2 F_\kappa(c_1,c_2)}\right| \end{aligned}$$

Here, the expectation is over $(\Lambda, W, G) \sim \mu \otimes \mathcal{N}(0,1) =: \mathcal{Q}$ with μ, $\mathbf{prox}_s(\cdot)$ and $F_\kappa(\cdot,\cdot)$ defined as in (2.4), (3.10) and (2.7) respectively.

Note that Λ denotes both the random variable in (3.11) and the covariance matrix in Assumption 1. Such overload of notations 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 Section 5. The max-ℓ_2-margin has received considerable attention in the past [67, 82, 43], however, (3.11) differs significantly from the equation

system considered in case of the ℓ_2 geometry. This is natural, due to the intrinsic differences between the ℓ_2 and ℓ_1 geometries, and this also leads to significant additional technical callenges in our setting (Section 5). 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 MLE, LRT [87, 98] and convex regularized estimators [86, 78] for logistic regression.

Uniqueness. Theorems 3.1-3.3 expressed our limiting results in terms of the solution to the system (3.11). It is, therefore, crucial to establish that the solution will indeed be unique. To this end, introduce the constants

$$\zeta = \left(\mathop{\mathbf{E}}_{(\Lambda,W)\sim\mu} |\Lambda^{-1/2} W| \right)^{-1} \qquad \text{and} \qquad \omega = \left(\mathop{\mathbf{E}}_{(\Lambda,W)\sim\mu} \left[(1-\zeta^2\Lambda^{-1})^2 W^2 \right] \right)^2 . \tag{3.12}$$

In a similar spirit as in [67], 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

$$\begin{aligned}
\psi_+(\kappa) &= \begin{cases} 0 & \text{if } \partial_1 F_\kappa(\zeta,0) > 0 \\ \partial_2^2 F_\kappa(-\zeta,0) - \omega^2 \partial_1^2 F_\kappa(-\zeta,0) & \text{if otherwise} \end{cases}, \\
\psi_-(\kappa) &= \begin{cases} 0 & \text{if } \partial_1 F_\kappa(-\zeta,0) > 0 \\ \partial_2^2 F_\kappa(\zeta,0) - \omega^2 \partial_1^2 F_\kappa(\zeta,0) & \text{if otherwise} \end{cases}, \\
\psi^{\downarrow}(\kappa) &= \max\{\psi^\star(\rho,f), \psi_+(\kappa), \psi_-(\kappa)\},
\end{aligned} \tag{3.13}$$

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

Proposition 3.1. *For any (ψ,κ) pair satisfying $\psi > \psi^{\downarrow}(\kappa)$, under the Assumptions 1-3, the system of equations (3.11) admits a unique solution $(c_1,c_2,s) \in \mathbb{R}\times\mathbb{R}_{>0}\times\mathbb{R}_{>0}$.*

Note that due to the definition of $\psi^{\downarrow}(\kappa)$, the region $\{(\psi,\kappa) : \psi > \psi^{\downarrow}(\kappa)\}$ contains $\{(\psi,\kappa) : T(\psi,\kappa) = 0\}$. It can be shown using arguments similar to [67, Section B.5], that the function $(\psi,\kappa) \to T(\psi,\kappa)$ is continuous on its domain, and on fixing $\psi > 0$, $T(\psi,\cdot)$ is strictly increasing with respect to κ with $\lim_{\kappa\to\infty} T(\psi,\kappa) = \infty$. Together, these ensure that (3.3) is well-defined, and that $c_1^\star, c_2^\star, s^\star$ are unique.

3.3 Boosting in high dimensions

We turn our attention to the *Boosting Algorithm* described in Section 2, Eqn. (2.9). The path of boosting iterates was studied in infinite time and infinitesimal stepsize in [75, 97]. Here, we establish a sharp analysis of the number of iterations necessary for the AdaBoost iterates to approximately maximize the ℓ_1-margin with arbitrary accuracy.

Theorem 3.4. *Under the assumptions of Theorem 3.1, with a suitably chosen learning rate (specified in Cor. 5.1), 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*

$$t \geq T_\epsilon(n) \quad \textit{with} \lim_{n\to\infty} \frac{T_\epsilon(n)}{n\log^2 n} \overset{\text{a.s.}}{=} \frac{12\psi}{\kappa_\star^2(\psi,\rho,\mu)} \epsilon^{-2} , \tag{3.14}$$

the solution $\hat\theta^t/\|\hat\theta^t\|_1$ forms $(1-\epsilon)$-approximation to the Min-ℓ_1-Interpolated Classifier, that is, almost surely,

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

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 notably complements the classical margin upper bounds [81, 53]. Thus, Theorem 3.4 reinforces a crucial conclusion from Section 3.1—the max-ℓ_1-margin is the key quantity governing the generalization error of AdaBoost, and as emphasized earlier, this resolves a series of discussions around this topic [81, 15].

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 ϵ, the stopping time $T_\epsilon(n)$ has a sharp asymptotic characterization (even in terms of constants), which contributes an essential insight to the computational properties of AdaBoost. To see this, Figure 3 plots the scaled margin limit $\psi^{-1/2}\kappa_\star(\psi,\rho,\mu)$ as a function of ψ, in the setting of Figure 1. The increase in this (scaled) limit as a function of ψ, together with (3.14), directly 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-ℓ_1-margin is once again the central quantity, as elucidated by (3.14).

Remark 3.1. *A natural question may be posed at this point: does the max-ℓ_1-margin studied here, when appropriately scaled, differ significantly from the ℓ_2-margin [67]? Note that the rescaled ℓ_1-margin is always larger than the ℓ_2-margin, denoted by κ_{n,ℓ_2}, since*

$$\kappa_{n,\ell_2} \le \sqrt{p}\cdot\kappa_{n,\ell_1}\ , \qquad \text{where} \qquad \kappa_{n,\ell_2} := \max_{\|\theta\|_2\le 1}\min_{1\le i\le n} y_i x_i^\top\theta\ . \tag{3.15}$$

A comparison of Figure 3 with [67, Fig. 1] shows that the range for the ℓ_1-margin is roughly twice that for the ℓ_2 case, demonstrating that these behave differently, even after appropriate scaling.

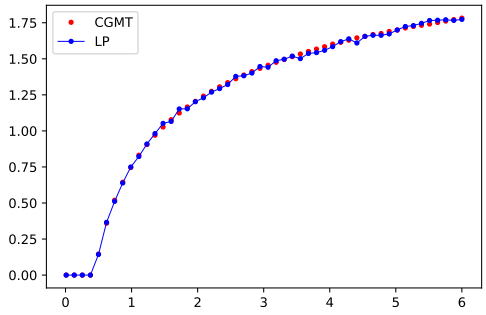

Figure 3: x-axis: varying ratio $\psi := p/n$. y-axis: $\kappa_\star/\sqrt{\psi}$.

Proportion of activated features for AdaBoost. The connection between the boosting solution and max-ℓ_1-margin naturally 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,*

$$S_0(p) := \#\left\{j \in [p] : \hat{\theta}_j^t \neq 0\right\} \ , \qquad \textit{where} \qquad \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{y_i x_i^\top \hat{\theta}^t \leq 0} = 0. \tag{3.16}$$

Under the assumptions of Theorem 3.4, $S_0(p)$, scaled appropriately, is asymptotically bounded by

$$\limsup_{p\to\infty} \frac{S_0(p)}{p\log^2 p} \leq \frac{12}{\kappa_\star^2(\psi,\rho,\mu)}, \ \ a.s. \tag{3.17}$$

This corollary provides specific insights into the geometry of the boosting solution, and formalizes the intuition that boosting must lead to sparse solutions. Note once again that the bound involves the max-ℓ_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 geometry of the boosting solution, beyond its foregoing roles in terms of generalization and optimization.

3.4 A new class of boosting algorithms

This section studies variants of AdaBoost that converge to the max-ℓ_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-ℓ_q-margin and the min-ℓ_q-norm interpolant beyond the $q = 1$ case. This complements the study of general ℓ_q constraints, that was initiated by [75]. To this end, define the max-ℓ_q-margin to be

$$\kappa_{n,\ell_q} := \max_{\|\theta\|_q \leq 1} \min_{1\leq i\leq n} \ y_i x_i^\top \theta \ , \tag{3.18}$$

and the corresponding min-ℓ_q-norm interpolant to be

$$\hat{\theta}_{n,\ell_q} \in \arg\min_{\theta} \ \|\theta\|_q, \ \text{ s.t. } y_i x_i^\top \theta \geq 1, \ 1 \leq i \leq n \ . \tag{3.19}$$

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 ℓ_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 θ_T.

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

Corollary 3.2 (Boosting Converges to max-ℓ_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 θ_T converge to the max-ℓ_q-margin Direction in the following sense: for any $0 < \epsilon < 1$,*

$$\kappa_{n,\ell_q} \geq \min_{i\in[n]} \frac{y_i x_i^\top \theta_T}{\|\theta_T\|_q} > \kappa_{n,\ell_q} \cdot (1-\epsilon), \tag{3.20}$$

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-ℓ_q-margin and the corresponding min-ℓ_q-norm interpolant admit analogous characterizations with a system of equations that differs from (3.11), with all else remaining the same. To introduce the equation system corresponding to general ℓ_q geometry, define the proximal mapping operator of the function $f_\lambda(t) = \lambda|t|^q$, for $\lambda > 0, q \geq 1$, to be

$$\mathbf{prox}_\lambda^{(q)}(t) := \arg\min_s \left\{ \lambda|s|^q + \frac{1}{2}(s-t)^2 \right\}. \tag{3.21}$$

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

$$c_1 = \langle \Lambda^{1/2} h^\star, W \rangle_{L_2(\mathcal{Q}_\infty)}, \qquad c_1^2 + c_2^2 = \|\Lambda^{1/2} h^\star\|_{L_2(\mathcal{Q}_\infty)}^2, \qquad \|h^\star\|_{L_q(\mathcal{Q})} = 1, \tag{3.22}$$

where $\mathcal{Q} = \mu \times \mathcal{N}(0,1)$. 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-ℓ_q-margin obeys,*

$$p^{\frac{1}{q}-\frac{1}{2}} \kappa_{n,\ell_q} \xrightarrow{\text{a.s.}} \kappa_\star^{(q)}(\psi,\rho,\mu), \tag{3.23}$$

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 (3.22). Simultaneously, the generalization error of the min-ℓ_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 (3.22), 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 $1 \le q \le 2$, with (3.11) swapped for (3.22). Once again, observe that the max-ℓ_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-ℓ_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).

Remark 3.2. *Note that Corollary 3.3 assumes the data is asymptotically linearly separable, that is, $\psi > \psi^\star(\rho, f)$. This threshold for separability 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 (3.23).*

3.5 Flexible Extensions

The theory presented so far provides precise characterizations of AdaBoost but relies, nonetheless, on the data generating process (2.1). This section explores relaxations of this assumption along two natural directions. (a) The model itself may be misspecified: we explore a common source of misspecification that occurs when the model misses a fraction of relevant variables. (b) The data generating process is different: we consider Gaussian mixture models commonly used to model classification problems. Studying AdaBoost and the max-ℓ_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 much broader class of data generation schemes.

3.5.1 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 Λ_x is a diagonal matrix. Suppose that y arises from the following conditional distribution

$$\mathbb{P}(y_i = +1 | \tilde{x}_i) = f(\tilde{x}_i^\top \theta_\star), \text{ with } \theta_\star := (\theta_{x,\star}^\top, \theta_{z,\star}^\top)^\top. \tag{3.24}$$

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

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

Consider that both components of $\theta_\star$, (3.24), contribute a non-trivial signal strength, in the sense that

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

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

$$\tilde{F}_\kappa(c_1, c_2) := \left(\mathbb{E}\left[(\kappa - c_1 Y Z_1 - c_2 Z_3)_+^2\right]\right)^{\frac{1}{2}}$$
$$\text{where} \quad \begin{cases} Z_3 \perp (Y, Z_1, Z_2) \\ Z_i \overset{i.i.d.}{\sim} \mathcal{N}(0,1),\ i = 1,2,3 \\ \mathbb{P}(Y = +1 | Z_1, Z_2) = 1 - \mathbb{P}(Y = -1 | Z_1, Z_2) = f(\rho \cdot Z_1 + \gamma \cdot Z_2) \end{cases} \tag{3.25}$$

Then the max-ℓ_1-margin and min-ℓ_1-norm interpolant, computed using the observed data $\{(x_i, y_i)\}_{i=1}^n$ obey the same limiting characterizations as in Theorems 3.1-3.3, with the system of equations remaining the same as in (3.11), but with $F_\kappa(c_1, c_2)$ substituted by the new function (3.25). Thus, the form of the equation system (3.11) remains unchanged, once we pin down the right analogue of $F_\kappa(c_1, c_2)$ in this new setting.

3.5.2 Gaussian Mixture Model

Consider the following Gaussian mixture model:

$$\mathbb{P}(y_i = +1) = 1 - \mathbb{P}(y_i = -1) = \upsilon \in (0,1)$$
$$x_i | y_i \sim \mathcal{N}(y_i \cdot \theta_\star, \Lambda),$$

where $\Lambda \in \mathbb{R}^{p \times p}$ is a diagonal matrix. Similar to Assumption 2, let $p(n)/n = \psi$ and denote

$$\frac{1}{p} \sum_{i=1}^{p} \delta_{(\lambda_i, \sqrt{p}\theta_\star^\top e_i)} \overset{W_2}{\Rightarrow} \mu. \tag{3.26}$$

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

$$\bar{F}_\kappa(c_1, c_2) := \left(\mathbb{E}\left[(\kappa - c_1 - c_2 Z)_+^2\right]\right)^{\frac{1}{2}} \text{ where } Z \sim \mathcal{N}(0,1). \tag{3.27}$$

Denote a triplet of random variables $(\Lambda, \Theta, G) \sim \mu \otimes \mathcal{N}(0,1) =: \mathcal{Q}$ with μ given by (3.26), and for any $\psi > 0$, define the following system of equations in variables $(c_1, c_2, s) \in \mathbb{R}^3$,

$$\begin{aligned} c_1 &= - \mathop{\mathbf{E}}_{(\Lambda,\Theta,G)\sim\mathcal{Q}} \left(\frac{\Lambda^{-1}\Theta \cdot \mathbf{prox}_s\left(\Lambda^{1/2} G + \psi^{-1/2} \partial_1 \bar{F}_\kappa(c_1, c_2)\Theta\right)}{\psi^{-1/2} c_2^{-1} \partial_2 F_\kappa(c_1, c_2)} \right) \\ c_2^2 &= \mathop{\mathbf{E}}_{(\Lambda,\Theta,G)\sim\mathcal{Q}} \left(\frac{\Lambda^{-1/2} \mathbf{prox}_s\left(\Lambda^{1/2} G + \psi^{-1/2} \partial_1 \bar{F}_\kappa(c_1, c_2)\Theta\right)}{\psi^{-1/2} c_2^{-1} \partial_2 \bar{F}_\kappa(c_1, c_2)} \right)^2 \\ 1 &= \mathop{\mathbf{E}}_{(\Lambda,\Theta,G)\sim\mathcal{Q}} \left| \frac{\Lambda^{-1} \mathbf{prox}_s\left(\Lambda^{1/2} G + \psi^{-1/2} \partial_1 \bar{F}_\kappa(c_1, c_2)\Theta\right)}{\psi^{-1/2} c_2^{-1} \partial_2 \bar{F}_\kappa(c_1, c_2)} \right|. \end{aligned} \tag{3.28}$$

Then the max-ℓ_1-margin and min-ℓ_1-norm interpolant obey the limiting characterizations from Theorems 3.1-3.3, with the system of equations given by (3.28), and $F_\kappa(c_1, c_2)$ substituted by (3.27).

In both the settings of Section 3.5.1 and here, AdaBoost satisfies Theorem 3.4 with the respective limiting characterization of the max-ℓ_1-margin. A common theme across all these settings is that

the behavior of the margin and interpolant can be accurately characterized by a three-dimensional system of equations, the solution to which possesses precise physical meanings (see (3.7) and the discussion thereafter) that remain the same across the different data generation schemes. The form of the systems vary from one model to another; however, the principles underlying its origin, and the key proof steps remain essentially the same (Section 5). Once again, this is the power of our theoretical analysis in the ℓ_1 case: we introduce a new uniform deviation argument with sufficient generality so that our proof can be adapted across a host of different modeling schemes, as illustrated through this section.

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 [39, 40], there has been a vast and expansive literature on Boosting. [13] 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 [50]. Margin-based analyses were furthered in [73, 54, 72, 74]. For analysis of boosting algorithms based on smooth margin functions, see [77] and the references cited therein. Consistency properties were extensively studied in [62, 64, 63, 11]. 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 [41]. Subsequently, [42] developed a general gradient boosting framework. The rate of convergence of regularized boosting classifiers was explored in [12], where the authors uncovered that some versions of boosting work especially well in high-dimensional logistic additive models. ℓ_2-boosting, sparse boosting, twin boosting, and their properties in high dimensions were extensively studied in [20, 18, 21, 17, 19]. We remark that our setting is different in nature from this high-dimensional Boosting literature, where a notion of sparsity (often in ℓ_1 geometry) is typically assumed on the unknown parameter $\theta_\star$. On the contrary, the ℓ_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 [68]. Robust versions of boosting were proposed and extensively explored in [56]. In recent times, [36] 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 [44, 45], and is obtained by extending Gordon's inequalities with the presence of convexity. The idea of merging these seemingly disparate threads dates back to [83, 84, 85], where it was used to analyze the performance of the constrained LASSO in high signal-to-noise ratio regimes. The seminal works [91, 90, 92] 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, [67] studied the asymptotic properties of the max-ℓ_2-margin in binary classification settings, building upon CGMT-based techniques, and furthered the work by [43]. 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, e.g. to characterize the

performance of the SLOPE estimator in sparse linear regression [49], to study high-dimensional regularized estimators in logistic regression [78], 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 entirely problem-specific and is usually rather challenging in most high-dimensional settings, often requiring the development of non-trivial probabilistic analysis (see Section 5 for specific details in our case).

Min-norm interpolation. This paper investigates the min-ℓ_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 [8, 9, 57, 10, 48, 6, 59, 58, 22] for the regression problem, and [67, 29, 24] 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 [95, 8, 57]. To the best of our knowledge, the current paper is the first to provide sharp statistical results for interpolated classifiers induced by the ℓ_1 geometry (rather than the ℓ_2), which has been argued to be a more suitable geometry [4, 47, 32, 25, 3] 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 Proof Sketch and Technical Contributions

The proofs of Theorems 3.1 and 3.2 rely on the *Convex Gaussian Min-Max Theorem* (CGMT) [91, 90], which is a refinement of Gordon's classical Gaussian comparison inequality [45]. Our analysis is partially influenced by the seminal work of [67], which characterized the max-ℓ_2-margin using CGMT-based techniques. However, characterizing the asymptotics for the ℓ_1 case requires establishing a novel and stronger form of a uniform deviation argument (outlined in Step 3 below); this relies on a key *self-normalizing* property of F_κ, which might be of standalone interest (we establish this in Lemma 5.1). Additionally, our analysis is general and extendable to the max-ℓ_q-margin case with $1 \le q \le 2$. Below, we provide a sketch of the main proof ideas.

5.1 Proofs of Theorems 3.1 and 3.2

Step 1: A basic reduction. To begin with, define

$$\begin{aligned}\xi_{\psi,\kappa}^{(n,p)} &:= \min_{\|\theta\|_1 \le \sqrt{p}} \max_{\|\lambda\|_2 \le 1, \lambda \ge 0} \frac{1}{\sqrt{p}} \lambda^T (\kappa \mathbf{1} - (y \odot X)\theta) \\ &= \min_{\|\theta\|_1 \le \sqrt{p}} \frac{1}{\sqrt{p}} \|(\kappa \mathbf{1} - (y \odot X)\theta)_+\|_2 \ .\end{aligned} \tag{5.1}$$

It is not hard to see that

$$\begin{aligned}&\xi_{\psi,\kappa}^{(n,p)} = 0, \ \text{ if and only if } \ \kappa \le p^{1/2} \cdot \kappa_{\ell_1}\left(\{x_i, y_i\}_{i=1}^n\right) \ , \\ &\xi_{\psi,\kappa}^{(n,p)} > 0, \ \text{ if and only if } \ \kappa > p^{1/2} \cdot \kappa_{\ell_1}\left(\{x_i, y_i\}_{i=1}^n\right) \ .\end{aligned} \tag{5.2}$$

Thus, to study the rescaled max-ℓ_1-margin, it suffices to examine the value of $\xi^{(n,p)}_{\psi,\kappa}$.

Now, defining $z_i := \Lambda^{-1/2} x_i \ \forall i \in [n]$, where Λ is the covariance matrix, we may express

$$x_i^\top \theta_\star = z_i^\top \Lambda^{1/2} \theta_\star = \rho_p \cdot z_i^\top w, \ \text{ where } w := \Lambda^{1/2}\theta_\star / \|\Lambda^{1/2}\theta_\star\| \ . \tag{5.3}$$

Using the fact that $y \odot X = (y \odot Z)\Lambda^{1/2} \overset{d}{=} ((y \odot z)w^\top + Z\Pi_{w^\perp})\Lambda^{1/2}$, where $z \in \mathbb{R}^n, Z \in \mathbb{R}^{n\times p}$ are independent of each other, each containing independent standard Gaussian entries, Eqn. (5.1) then reduces to

$$\xi^{(n,p)}_{\psi,\kappa}(z,Z) := \min_{\|\theta\|_1 \le \sqrt{p}} \max_{\|\lambda\|_2 \le 1, \lambda \ge 0} \frac{1}{\sqrt{p}} \lambda^T \left(\kappa \mathbf{1} - (y \odot z)\langle w, \Lambda^{1/2}\theta\rangle\right) - \frac{1}{\sqrt{p}} \lambda^T Z \Pi_{w^\perp}(\Lambda^{1/2}\theta) \ . \tag{5.4}$$

Remark 5.1. *The rescaling by $\sqrt{p}$ is required to ensure a well-defined limit for the max-ℓ_1-margin (in general, a rescaling by $p^{1/q-1/2}$ is required for general ℓ_q margin, as evidenced via Corollary 3.23, and this immediately shows that no rescaling is required for the ℓ_2 case [67]).*

Step 2: Reduction to Gordon's problem. Due to the min-max form of (5.4), one can use Gordon's Gaussian comparison inequality [91, 90, 45] to further simplify the problem. To this end, introduce the following "de-coupled" optimization problem

$$\begin{aligned}
&\hat{\xi}^{(n,p)}_{\psi,\kappa}(z,\tilde{z},g) \\
&:= \min_{\|\theta\|_1 \le \sqrt{p}} \max_{\|\lambda\|_2 \le 1, \lambda \ge 0} \frac{1}{\sqrt{p}} \lambda^T \left(\kappa \mathbf{1} - (y \odot z)\langle w, \Lambda^{1/2}\theta\rangle - \tilde{z}\|\Pi_{w^\perp}(\Lambda^{1/2}\theta)\|_2\right) + \frac{1}{\sqrt{p}}\|\lambda\|_2 \langle g, \Pi_{w^\perp}(\Lambda^{1/2}\theta)\rangle \\
&= \left[\min_{\|\theta\|_1 \le \sqrt{p}} \frac{1}{\sqrt{p}} \left\| \left(\kappa \mathbf{1} - (y \odot z)\langle w, \Lambda^{1/2}\theta\rangle - \tilde{z}\|\Pi_{w^\perp}(\Lambda^{1/2}\theta)\|_2\right)_+ \right\|_2 + \frac{1}{\sqrt{p}} \left\langle \Pi_{w^\perp}(g), \Lambda^{1/2}\theta \right\rangle \right]_+ ,
\end{aligned} \tag{5.5}$$

where $z, \tilde{z} \in \mathbb{R}^n$ and $g \in \mathbb{R}^p$ are independent isotropic Gaussian vectors. By CGMT [91, Theorem 3] (see Theorem A.1 in the Appendix), we have

$$\mathbb{P}\left(\xi^{(n,p)}_{\psi,\kappa}(z,Z) \le t | y, z\right) \le 2\mathbb{P}\left(\hat{\xi}^{(n,p)}_{\psi,\kappa}(z,\tilde{z},g) \le t | y, z\right) \tag{5.6}$$

$$\mathbb{P}\left(\xi^{(n,p)}_{\psi,\kappa}(z,Z) \ge t | y, z\right) \le 2\mathbb{P}\left(\hat{\xi}^{(n,p)}_{\psi,\kappa}(z,\tilde{z},g) \ge t | y, z\right). \tag{5.7}$$

Marginalizing over y and z, this suggests that it suffices to study (5.5).

Step 3: The key step—large n, p limit, new uniform deviation result.

Recall the function $F_\kappa(\cdot,\cdot)$ from (2.7), and define the empirical version

$$\widehat{F}_\kappa(c_1, c_2) := \left(\widehat{\mathbf{E}}_n[(\kappa - c_1 Y Z_1 - c_2 Z_2)_+^2]\right)^{1/2} \ , \tag{5.8}$$

where $\widehat{\mathbf{E}}_n$ means that the expectation over Y, Z_1, Z_2 is taken with respect to the empirical distribution of $\{(Y_i, Z_{1,i}, Z_{2,i})\}_{i=1}^n$, with entries $(Y_i, Z_{1,i}, Z_{2,i})$ arising from the joint distribution specified in (2.7). Then with $\lambda = \mathrm{diag}(\Lambda)$ denoting the vectorized Λ, we can express $\hat{\xi}^{(n,p)}_{\psi,\kappa}(z,\tilde{z},g)$ as the positive part of the following expression

$$\hat{\xi}^{(n,p)}_{\psi,\kappa}(\lambda, w, g) := \min_{\|\theta\|_1 \le \sqrt{p}} \left[\psi^{-1/2} \widehat{F}_\kappa\left(\langle w, \Lambda^{1/2}\theta\rangle, \|\Pi_{w^\perp}(\Lambda^{1/2}\theta)\|_2\right) + \frac{1}{\sqrt{p}} \left\langle \Pi_{w^\perp}(g), \Lambda^{1/2}\theta \right\rangle\right] \ . \tag{5.9}$$

Note that $\hat{\xi}^{(n,p)}_{\psi,\kappa}(\lambda, w, g)$ is a random quantity, here we denote λ, w, g as arguments to make explicit the dependence.

We seek to study (5.9) in the large sample and feature limits $n, p \to \infty$ with $p/n \to \psi$. On taking limits naively, one can reach the following infinite-dimensional convex problem,

$$\tilde{\xi}^{(\infty,\infty)}_{\psi,\kappa}(\Lambda, W, G) := \min_{\|h\|_{L_1(\mathcal{Q})} \le 1} \left[\psi^{-1/2} F_\kappa \left(\langle W, \Lambda^{1/2} h \rangle_{L_2(\mathcal{Q})}, \|\Pi_{W^\perp}(\Lambda^{1/2} h)\|_{L_2(\mathcal{Q})} \right) + \left\langle \Pi_{W^\perp}(G), \Lambda^{1/2} h \right\rangle_{L_2(\mathcal{Q})} \right]. \tag{5.10}$$

Here, the optimization variable is the set of function $\{h : \mathbb{R}^3 \to \mathbb{R}, h \in \mathcal{L}^2(\mathcal{Q})\}$, where $\mathcal{Q} = \mu \otimes \mathcal{N}(0,1)$ with μ defined as in (2.4).

Proposition A.1 rigorously proves that the empirical optimization problem $\hat{\xi}^{(n,p)}_{\psi,\kappa}(\lambda, w, g)$ converges to the infinite dimensional problem $\tilde{\xi}^{(\infty,\infty)}_{\psi,\kappa}(\Lambda, W, G)$, almost surely, that is,

$$\lim_{n\to\infty, p(n)/n=\psi} \hat{\xi}^{(n,p)}_{\psi,\kappa}(\lambda, w, g) \overset{\text{a.s.}}{=} \tilde{\xi}^{(\infty,\infty)}_{\psi,\kappa}(\Lambda, W, G). \tag{5.11}$$

We provide an outline of the proof below, deferring the details to Section A.2.

Our *technical innovation* lies in the development of (5.11), which requires establishing a uniform deviation bound over an unbounded region. To describe further, observe that $\hat{\xi}^{(n,p)}_{\psi,\kappa}(\lambda, w, g)$ involves $\hat{F}_\kappa$ evaluated at the points $c_1 = \langle w, \Lambda^{1/2}\theta \rangle$ and $c_2 = \|\Pi_{w^\perp}(\Lambda^{1/2}\theta)\|_2$. It is clear that both under the ℓ_2-constraint $\|\theta\|_2 \le 1$ (the setting of [67]) and the ℓ_1-constraint $\|\theta\|_1 \le \sqrt{p}$ (our setting), c_1 is bounded in the sense $|c_1| \le M$ for all $p(n), n$ and some constant $M > 0$; for the ℓ_1 case, this follows by noting that

$$|\langle w, \Lambda^{1/2}\theta \rangle| \le \frac{1}{c} \cdot \|w\|_\infty \|\theta\|_1 = \frac{1}{c} \cdot \|\bar{w}\|_\infty / \sqrt{p} \cdot \|\theta\|_1 \le C'/c,$$

by Assumption 3. Turning to the second variable c_2, we see that under our ℓ_1-constraint, c_2 may potentially grow as $\sqrt{p}$ whereas it remains bounded when the ℓ_2-norm of θ is bounded. Naturally, the unbounded region for c_2 creates significant challenges in establishing (5.11) in our setting. Naive covering arguments to establish the aforementioned uniform deviation for $c_2 \in [0, \infty)$ fails to deliver sharp results. To overcome this technical challenge, we discover a key self-normalization property of the partial derivatives of F_κ (Appendix A.2), utilizing the structure of this function, and prove the following.

Lemma 5.1 (Self-normalization and uniform deviation). *For $i = 1, 2$, with probability at least $1 - n^{-2}$,*

$$\sup_{|c_1| \le M, c_2 > 0} |\partial_i \widehat{F}_\kappa(c_1, c_2) - \partial_i F_\kappa(c_1, c_2)| \le C \cdot \frac{\log n}{\sqrt{n}}, \tag{5.12}$$

where C is a constant that does not depend on n.

Our proof proceeds as follows: (a) The first and key step is to establish Lemma 5.1. (b) Thereafter, we establish that the "empirical fixed point (fp) equations" obtained by analyzing the KKT conditions for (5.9)[1] converge uniformly (over an unbounded region for c_2), to the corresponding "fp equations obtained from the KKT conditions for (5.10)".[2] The convergence here

[1] This finite n, p problem is not convex in θ, the KKT conditions are merely necessary conditions in this case.

[2] The KKT conditions are both necessary and sufficient in this case. See Appendix A.2 for details.

is in the sense of (A.12). The analysis uses the key Lemma 5.1. See Step 4 for description of these KKT equations. (c) Leveraging (b), we show that any solution $(\hat{c}_1, \hat{c}_2, \hat{s})$ of the empirical fp equations converges to the unique solution $(c_1^\star, c_2^\star, s^\star)$ of the fp equations from (5.10). See Appendix A.3 for uniqueness of the solution. (d) Now, (5.9) can be expressed as functions of $\hat{s}$ and $\hat{F}_\kappa, \partial_i \hat{F}_\kappa, i = 1, 2$, evaluated at $(\hat{c}_1, \hat{c}_2)$, and similarly, for (5.10) with $s^\star$ and $F_\kappa, \partial_i F_\kappa, i = 1, 2$ evaluated at $(c_1^\star, c_2^\star)$. Given (c), we have proved that $(\hat{c}_1, \hat{c}_2, \hat{s})$ will be bounded for sufficiently large n, and therefore, uniform deviation bounds for $|\hat{F}_\kappa - F_\kappa|$ can also be established. This series of arguments enables us to establish (5.11), under a potentially complicated ℓ_1 geometry. A critical, and perhaps surprising, consequence of our uniform deviation results is a localization property: any optimizer of (5.9) possesses finite ℓ_2-norm.

Step 4: Fixed point equations and final step.

By standard analysis arguments (see Appendix A.3), the KKT conditions for the optimization problem (5.10) can be expressed as

$$\begin{aligned} &\Pi_{W^\perp}(G) + \psi^{-1/2}\Big[\partial_1 F_\kappa(c_1, c_2) W + c_2^{-1} \partial_2 F_\kappa(c_1, c_2)(\Lambda^{1/2} h - c_1 W)\Big] + s \cdot \Lambda^{-1/2} \partial \|h\|_{L_1(\mathcal{Q})} = 0, \\ &\text{and} \quad \|h\|_{L_1(\mathcal{Q})} = 1, \quad \text{where} \quad c_1 := \langle \Lambda^{1/2} h, W \rangle_{L_2(\mathcal{Q})}, \quad c_2 := \|\Pi_{W^\perp}(\Lambda^{1/2} h)\|_{L_2(\mathcal{Q})}. \end{aligned} \tag{5.13}$$

From properties of the proximal mapping operator, the KKT conditions suggest that the solution must satisfy (see Appendix A.3 for uniqueness of solution)

$$h = -\frac{\Lambda^{-1} \, \mathbf{prox}_s\Big(\Lambda^{1/2} G + \psi^{-1/2}[\partial_1 F_\kappa(c_1, c_2) - c_1 c_2^{-1} \partial_2 F_\kappa(c_1, c_2)] \Lambda^{1/2} W\Big)}{\psi^{-1/2} c_2^{-1} \partial_2 F_\kappa(c_1, c_2)}. \tag{5.14}$$

Plugging this in the three equations displayed in (5.13), leads to the "fp equations … for (5.10)", referred to in Step 3, which is the exact same as the equation system (3.11), thus explaining the origin of the system. A similar analysis for (5.9) leads to the "empirical fp equations" referred to in Step 3 (see (A.11) for the specific form). Finally, Corollary A.1 shows that $\tilde{\xi}_{\psi,\kappa}^{(\infty,\infty)}(\Lambda, W, G) = T(\psi, \kappa)$; together with (5.2) and (5.11), this completes the proof.

Note that (5.14) clarifies how $s^\star$ from Section 3.1 (see in particular, the discussion following 3.10) corresponds to Lagrange multipliers induced by the ℓ_1 constraint in(5.13).

5.2 Proofs of Theorems 3.4 and Corollary 3.1

[97] employs a re-scaling technique to establish that Boosting with infinitesimal stepsize agrees with the *min-ℓ_1-norm* direction asymptotically. Since we care about the actual number of iterations in the Boosting algorithm (which translates to the number of selected features), here we provide a simple yet general analysis of Boosting as a special instance of Mirror Descent, in conjunction with the re-scaling technique [97] and the shrinkage technique [89]. Our analysis provides a sharp upper bound on the number of iterations of the algorithm, and is similar in spirit to [27], but with different executions. One benefit of our analysis is that it is easily generalizable to a variant of boosting algorithm that maximizes ℓ_q margin with $q \geq 1$, which is new to the literature.

Proposition 5.1 (Boosting as mirror descent)**.** *Consider the* Boosting Algorithm *stated in Section 2 Eqn. 2.9. Assume that $|X_{ij}| \leq M$ for $i \in [n], j \in [p]$. Consider the learning rate $\alpha_t(\beta) = \beta \cdot \eta_t^\top Z v_{t+1}$, with $\beta = 1/M^2$. When*

$$T \geq \frac{2M^2}{\kappa_{n,\ell_1}^2} \log \frac{ne}{\epsilon}, \tag{5.15}$$

the Boosting Algorithm *iterates* θ_T *will satisfy* $\sum_{i\in[n]} 1_{x_i^\top \theta_T \leq 0} \leq \epsilon$.

Corollary 5.1 (Boosting converges to max-ℓ_1-margin direction)**.** *Consider the general 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* θ_T *converge to the max-*ℓ_1*-margin Direction in the following sense: for any* $0 < \epsilon < 1$,

$$\kappa_{n,\ell_1} \geq \min_{i\in[n]} \frac{y_i x_i^\top \theta_T}{\|\theta_T\|_1} > \kappa_{n,\ell_1} \cdot (1-\epsilon), \tag{5.16}$$

where $T \geq \log(1.01ne) \cdot \frac{2M^2\epsilon^{-2}}{\kappa_{n,\ell_1}^2}$, *with* $\beta = \frac{\epsilon}{M^2}$.

To obtain Theorems 3.4 and Corollary 3.1, we choose $M(\delta) = \sqrt{(3+\delta)\log(np)}$ for arbitrarily small $\delta > 0$. Now, the entries X_{ij} are uniformly bounded above by M asymptotically almost surely, since $\mathbb{P}(\sup_{i\in[n],j\in[p]} |X_{ij}| \leq M(\delta)) \leq np\exp(-M^2(\delta)/2) = n^{-1-\delta}$ and $\sum_{n\geq 1} n^{-1-\delta} < \infty$. Plugging in $\epsilon = 0.99$ in Proposition 5.1, with the aforementioned M, establishes the almost sure result in Theorem 3.1. The constant 12 can be justified since $\lim_{\delta\to 0} 2M^2(\delta)/\log n = 12$.

6 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-ℓ_1-margin, the min-ℓ_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, 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-ℓ_q-margins for $q > 1$, are further analyzed.

We conclude with a few 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 ℓ_q geometry when $q \neq 2$, (b) to compare the variants of AdaBoost corresponding to different ℓ_q geometry and pin down the model parameters that determine which variant would be optimal, and (c) to complement such characterizations via data-driven schemes for estimating these parameters, which may be used to provide recommendations regarding algorithm choice to practitioners.

7 Acknowledgements

T. Liang wishes to thank Yoav Freund and other participants in the Learning Theory seminar at Google Research for constructive feedback that greatly improved the paper. T. Liang acknowledges

support from the George C. Tiao Fellowship. P. Sur was partially supported by the Center for Research on Computation and Society, Harvard John A. Paulson School of Engineering and Applied Sciences. P. Sur wishes to thank the organizers and participants of the Young Data Science Researcher Seminar, ETH Zurich, for constructive feedback.

References

[1] Adelin Albert and John A Anderson. On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 71(1):1–10, 1984.

[2] Luigi Ambrosio and Nicola Gigli. A user's guide to optimal transport. In *Modelling and optimisation of flows on networks*, pages 1–155. Springer, 2013.

[3] Ehsan Amid and Manfred K Warmuth. Winnowing with gradient descent. *Proceedings of Machine Learning Research vol*, 125:1–20, 2020.

[4] Francis Bach. Breaking the curse of dimensionality with convex neural networks. *The Journal of Machine Learning Research*, 18(1):629–681, 2017.

[5] Peter L Bartlett and Mikhail Traskin. Adaboost is consistent. *Journal of Machine Learning Research*, 8(Oct):2347–2368, 2007.

[6] Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *arXiv preprint arXiv:1906.11300*, 2019.

[7] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine learning and the bias-variance trade-off. *arXiv preprint arXiv:1812.11118*, 2018.

[8] Mikhail Belkin, Siyuan Ma, and Soumik Mandal. To understand deep learning we need to understand kernel learning. *arXiv preprint arXiv:1802.01396*, 2018.

[9] Mikhail Belkin, Alexander Rakhlin, and Alexandre B Tsybakov. Does data interpolation contradict statistical optimality? *arXiv preprint arXiv:1806.09471*, 2018.

[10] Mikhail Belkin, Daniel Hsu, and Ji Xu. Two models of double descent for weak features. *arXiv preprint arXiv:1903.07571*, 2019.

[11] Peter J Bickel, Ya'acov Ritov, and Alon Zakai. Some theory for generalized boosting algorithms. *Journal of Machine Learning Research*, 7(May):705–732, 2006.

[12] Gilles Blanchard, Gábor Lugosi, and Nicolas Vayatis. On the rate of convergence of regularized boosting classifiers. *Journal of Machine Learning Research*, 4(Oct):861–894, 2003.

[13] Leo Breiman. Arcing classifiers. *Annals of Statistics*, 26:123–40, 1996.

[14] Leo Breiman. Bias, variance, and arcing classifiers. Technical report, Tech. Rep. 460, Statistics Department, University of California, Berkeley . . . , 1996.

[15] Leo Breiman. Prediction games and arcing algorithms. *Neural computation*, 11(7):1493–1517, 1999.

[16] Leo Breiman. Population theory for boosting ensembles. *The Annals of Statistics*, 32(1):1–11, 2004.

[17] Peter Buhlmann. Boosting for high-dimensional linear models. *The Annals of Statistics*, 34(2): 559–583, 2006.

[18] Peter Bühlmann and Torsten Hothorn. Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science*, 22(4):477–505, 2007.

[19] Peter Bühlmann and Torsten Hothorn. Twin boosting: improved feature selection and prediction. *Statistics and Computing*, 20(2):119–138, 2010.

[20] Peter Bühlmann and Bin Yu. Boosting with the l 2 loss: regression and classification. *Journal of the American Statistical Association*, 98(462):324–339, 2003.

[21] Peter Bühlmann and Bin Yu. Sparse boosting. *Journal of Machine Learning Research*, 7(Jun): 1001–1024, 2006.

[22] Florentina Bunea, Seth Strimas-Mackey, and Marten Wegkamp. Interpolation under latent factor regression models. *arXiv preprint arXiv:2002.02525*, 2020.

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

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

[25] Lenaic Chizat and Francis Bach. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. *arXiv preprint arXiv:2002.04486*, 2020.

[26] Alexandra Chouldechova and Aaron Roth. The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810*, 2018.

[27] Michael Collins, Robert E Schapire, and Yoram Singer. Logistic regression, adaboost and bregman distances. *Machine Learning*, 48(1-3):253–285, 2002.

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

[29] Zeyu Deng, Abla Kammoun, and Christos Thrampoulidis. A Model of Double Descent for High-dimensional Binary Linear Classification. *arXiv:1911.05822 [cs, eess, stat]*, November 2019.

[30] Oussama Dhifallah, Christos Thrampoulidis, and Yue M Lu. Phase retrieval via polytope optimization: Geometry, phase transitions, and new algorithms. *arXiv preprint arXiv:1805.09555*, 2018.

[31] David Donoho and Andrea Montanari. High dimensional robust m-estimation: Asymptotic variance via approximate message passing. *Probability Theory and Related Fields*, 166(3-4): 935–969, 2016.

[32] Xialiang Dou and Tengyuan Liang. Training Neural Networks as Learning Data-adaptive Kernels: Provable Representation and Approximation Benefits. *Journal of the American Statistical Association*, pages 1–35, March 2020. ISSN 0162-1459, 1537-274X. doi: 10.1080/01621459.2020.1745812.

[33] Harris Drucker and Corinna Cortes. Boosting decision trees. In *Advances in neural information processing systems*, pages 479–485, 1996.

[34] Noureddine El Karoui. On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators. *Probability Theory and Related Fields*, 170(1-2):95–175, 2018.

[35] Noureddine El Karoui, Derek Bean, Peter J Bickel, Chinghway Lim, and Bin Yu. On robust regression with high-dimensional predictors. *Proceedings of the National Academy of Sciences*, 110(36):14557–14562, 2013.

[36] Robert M Freund, Paul Grigas, and Rahul Mazumder. A new perspective on boosting in linear regression via subgradient optimization and relatives. *The Annals of Statistics*, 45(6): 2328–2364, 2017.

[37] Yoav Freund. Boosting a weak learning algorithm by majority. *Information and computation*, 121(2):256–285, 1995.

[38] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. In *European conference on computational learning theory*, pages 23–37. Springer, 1995.

[39] Yoav Freund and Robert E Schapire. A desicion-theoretic generalization of on-line learning and an application to boosting. In *European conference on computational learning theory*, pages 23–37. Springer, 1995.

[40] Yoav Freund and Robert E Schapire. Experiments with a new boosting algorithm. In *icml*, volume 96, pages 148–156. Citeseer, 1996.

[41] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The annals of statistics*, 28(2):337–407, 2000.

[42] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[43] Elizabeth Gardner. The space of interactions in neural network models. *Journal of physics A: Mathematical and general*, 21(1):257, 1988.

[44] Yehoram Gordon. Some inequalities for gaussian processes and applications. *Israel Journal of Mathematics*, 50(4):265–289, 1985.

[45] Yehoram Gordon. On milman's inequality and random subspaces which escape through a mesh in $\mathbb{R}$ n. In *Geometric aspects of functional analysis*, pages 84–106. Springer, 1988.

[46] Adam J Grove and Dale Schuurmans. Boosting in the limit: Maximizing the margin of learned ensembles. In *AAAI/IAAI*, pages 692–699, 1998.

[47] Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. *arXiv preprint arXiv:1802.08246*, 2018.

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

[49] Hong Hu and Yue M Lu. Asymptotics and optimal designs of slope for sparse linear regression. In *2019 IEEE International Symposium on Information Theory (ISIT)*, pages 375–379. IEEE, 2019.

[50] Wenxin Jiang. Some theoretical aspects of boosting in the presence of noisy data. In *Proceedings of the Eighteenth International Conference on Machine Learning*. Citeseer, 2001.

[51] Wenxin Jiang. Process consistency for adaboost. *The Annals of Statistics*, 32(1):13–29, 2004.

[52] Jon Kleinberg and Sendhil Mullainathan. Simplicity creates inequity: implications for fairness, stereotypes, and interpretability. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, pages 807–808, 2019.

[53] Vladimir Koltchinskii and Dmitry Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *The Annals of Statistics*, 30(1):1–50, 2002.

[54] Vladimir Koltchinskii and Dmitry Panchenko. Complexities of convex combinations and bounding the generalization error in classification. *The Annals of Statistics*, 33(4):1455–1496, 2005.

[55] Emmanuel Lesaffre and Adelin Albert. Partial separation in logistic discrimination. *Journal of the Royal Statistical Society: Series B (Methodological)*, 51(1):109–116, 1989.

[56] Alexander Hanbo Li and Jelena Bradic. Boosting in the presence of outliers: adaptive classification with nonconvex loss functions. *Journal of the American Statistical Association*, 113(522): 660–674, 2018.

[57] Tengyuan Liang and Alexander Rakhlin. Just interpolate: Kernel "Ridgeless" regression can generalize. *The Annals of Statistics*, 48(3):1329–1347, July 2020. ISSN 0090-5364. doi: 10.1214/19-AOS1849.

[58] Tengyuan Liang and Hai Tran-Bach. Mehler's formula, branching process, and compositional kernels of deep neural networks. *arXiv preprint arXiv:2004.04767*, 2020.

[59] Tengyuan Liang, Alexander Rakhlin, and Xiyu Zhai. On the Multiple Descent of Minimum-Norm Interpolants and Restricted Lower Isometry of Kernels. *arXiv:1908.10292*, *COLT 2020*, *to appear*, 2019.

[60] Zachary C Lipton. The mythos of model interpretability. *Queue*, 16(3):31–57, 2018.

[61] Gábor Lugosi and Nicolas Vayatis. On the Bayes-risk consistency of regularized boosting methods. *Annals of Statistics*, 32(1):30–55, February 2004. ISSN 0090-5364, 2168-8966. doi: 10.1214/aos/1079120129.

[62] Shie Mannor and Ron Meir. Geometric bounds for generalization in boosting. In *International Conference on Computational Learning Theory*, pages 461–472. Springer, 2001.

[63] Shie Mannor and Ron Meir. On the existence of linear weak learners and applications to boosting. *Machine Learning*, 48(1-3):219–251, 2002.

[64] Shie Mannor, Ron Meir, and Tong Zhang. The consistency of greedy algorithms for classification. In *International Conference on Computational Learning Theory*, pages 319–333. Springer, 2002.

[65] Llew Mason, Jonathan Baxter, Peter L Bartlett, and Marcus R Frean. Boosting algorithms as gradient descent. In *Advances in neural information processing systems*, pages 512–518, 2000.

[66] Andrea Montanari. Mean field asymptotics in high-dimensional statistics: From exact results to efficient algorithms. In *Proceedings of the International Congress of Mathematicians, World Scientific*, pages 2957–2980. World Scientific, 2018.

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

[68] Indraneel Mukherjee, Cynthia Rudin, and Robert E Schapire. The rate of convergence of adaboost. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 537–558, 2011.

[69] Ashkan Panahi and Babak Hassibi. A universal analysis of large-scale regularized least squares solutions. In *Advances in Neural Information Processing Systems*, pages 3381–3390, 2017.

[70] Neal Parikh and Stephen Boyd. Proximal algorithms. *Foundations and Trends® in Optimization*, 1(3):127–239, 2014.

[71] JR Quinlan. 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, 1996.

[72] Gunnar Rätsch and Manfred K Warmuth. Efficient margin maximizing with boosting. *Journal of Machine Learning Research*, 6(Dec):2131–2152, 2005.

[73] Gunnar Rätsch, Takashi Onoda, and K-R Müller. Soft margins for adaboost. *Machine learning*, 42(3):287–320, 2001.

[74] Lev Reyzin and Robert E Schapire. How boosting the margin can also boost classifier complexity. In *Proceedings of the 23rd international conference on Machine learning*, pages 753–760, 2006.

[75] Saharon Rosset, Ji Zhu, and Trevor Hastie. Boosting as a regularized path to a maximum margin classifier. *Journal of Machine Learning Research*, 5(Aug):941–973, 2004.

[76] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

[77] Cynthia Rudin, Robert E Schapire, and Ingrid Daubechies. Analysis of boosting algorithms using the smooth margin function. *The Annals of Statistics*, 35(6):2723–2768, 2007.

[78] Fariborz Salehi, Ehsan Abbasi, and Babak Hassibi. The impact of regularization on high-dimensional logistic regression. In *Advances in Neural Information Processing Systems*, pages 11982–11992, 2019.

[79] Thomas J Santner and Diane E Duffy. A note on a. albert and ja anderson's conditions for the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 73(3): 755–758, 1986.

[80] Robert E Schapire. The strength of weak learnability. *Machine learning*, 5(2):197–227, 1990.

[81] Robert E Schapire, Yoav Freund, Peter Bartlett, and Wee Sun Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. *The annals of statistics*, 26(5):1651–1686, 1998.

[82] Mariya Shcherbina and Brunello Tirozzi. Rigorous solution of the gardner problem. *Communications in mathematical physics*, 234(3):383–422, 2003.

[83] Mihailo Stojnic. A framework to characterize performance of lasso algorithms. *arXiv preprint arXiv:1303.7291*, 2013.

[84] Mihailo Stojnic. Meshes that trap random subspaces. *arXiv preprint arXiv:1304.0003*, 2013.

[85] Mihailo Stojnic. Upper-bounding l1-optimization weak thresholds. *available at arXiv*, 2013.

[86] Pragya Sur. A modern maximum-likelihood theory for high-dimensional logistic regression. purl.stanford.edu/jw604jq1260, *Ph.D. thesis, Stanford University*, 2019.

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

[88] Pragya Sur, Yuxin Chen, and Emmanuel J Candès. The likelihood ratio test in high-dimensional logistic regression is asymptotically a rescaled chi-square. *Probability theory and related fields*, 175(1-2):487–558, 2019.

[89] Matus Telgarsky. Margins, shrinkage, and boosting. *arXiv preprint arXiv:1303.4172*, 2013.

[90] Christos Thrampoulidis, Samet Oymak, and Babak Hassibi. The gaussian min-max theorem in the presence of convexity. *arXiv preprint arXiv:1408.4837*, 2014.

[91] Christos Thrampoulidis, Samet Oymak, and Babak Hassibi. Regularized linear regression: A precise analysis of the estimation error. In *Conference on Learning Theory*, pages 1683–1709, 2015.

[92] Christos Thrampoulidis, Ehsan Abbasi, and Babak Hassibi. Precise error analysis of regularized m-estimators in high dimensions. *IEEE Transactions on Information Theory*, 64(8): 5592–5628, 2018.

[93] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

[94] Adrian Weller. Transparency: motivations and challenges. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pages 23–40. Springer, 2019.

[95] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

[96] Tong Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics*, pages 56–85, 2004.

[97] Tong Zhang and Bin Yu. Boosting with early stopping: Convergence and consistency. *The Annals of Statistics*, 33(4):1538–1579, 2005.

[98] Qian Zhao, Pragya Sur, and Emmanuel J Candes. The asymptotic distribution of the mle in high-dimensional logistic models: Arbitrary covariance. *arXiv preprint arXiv:2001.09351*, 2020.

A Technical Proofs

A.1 The Convex Gaussian Min-Max Theorem

For the convenience of the readers, we state Gordon's comparison inequality [45] below [91, Theorem 4]. We state the form mentioned in [67, Theorem 2].

Theorem A.1. *Let $\Omega_1 \subset \mathbb{R}^n, \Omega_2 \subset \mathbb{R}^p$ be two compact sets and let $U : \Omega_1 \times \Omega_2 \to \mathbb{R}$ be a continuous function. Let $Z = (Z_{i,j}) \in \mathbb{R}^{n\times p}, g \sim \mathcal{N}(0, I_n)$ and $h \sim \mathcal{N}(0, I_p)$ be independent vectors and matrices with standard Gaussian entries. Define*

$$V_1(Z) = \min_{w_1\in\Omega_1} \max_{w_2\in\Omega_2} w_1^\top Z w_2 + U(w_1, w_2) \ ,$$
$$V_2(g,h) = \min_{w_1\in\Omega_1} \max_{w_2\in\Omega_2} \|w_2\| g^\top w_1 + \|w_1\| h^\top w_2 + U(w_1, w_2) \ .$$

Then

1. *For all $t \in \mathbb{R}$,*
$$\mathbb{P}(V_1(Z) \le t) \le 2\mathbb{P}(V_2(g,h) \le t) \ .$$
2. *Suppose Ω_1 and Ω_2 are both convex, and U is convex-concave in (w_1, w_2). Then, for all $t \in \mathbb{R}$,*
$$\mathbb{P}(V_1(Z) \ge t) \le 2\mathbb{P}(V_2(g,h) \ge t) \ .$$

A.2 Large n, p Limit: New Uniform Convergence Results

Let $g \in \mathbb{R}^n$ be such that $g_i \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$. Recall the definitions of λ_j, w_j from Assumption 1 and (5.3) respectively, and denote the empirical distribution of $\{(\lambda_j, \sqrt{p}w_j, g_j)\}_{i=1}^p$ by $\mathcal{Q}_p$, that is,

$$\mathcal{Q}_p = \frac{1}{p}\sum_{j=1}^p \delta_{(\lambda_j, \sqrt{p}w_j, g_j)}. \tag{A.1}$$

Simultaneously, let $\mathcal{Q}_\infty = \mathcal{Q}$ from Definition 1, that is, $\mathcal{Q}_\infty = \mu \otimes \mathcal{N}(0,1)$. Define the functions $V_1^{(\infty,\infty)}(\cdot,\cdot,\cdot), V_2^{(\infty,\infty)}(\cdot,\cdot,\cdot), V_3^{(\infty,\infty)}(\cdot,\cdot,\cdot) : \mathbb{R}^3 \to \mathbb{R}$ as follows

$$\begin{aligned}
&V_1^{(\infty,\infty)}(c_1,c_2,s) := \\
&c_1 + \mathop{\mathbf{E}}_{(\Lambda,W,G)\sim\mathcal{Q}_\infty}\left(\frac{\Lambda^{-1/2}W \cdot \mathbf{prox}_s\left(\Lambda^{1/2}\Pi_{W^\perp}(G) + \psi^{-1/2}[\partial_1 F_\kappa(c_1,c_2) - c_1c_2^{-1}\partial_2 F_\kappa(c_1,c_2)]\Lambda^{1/2}W\right)}{\psi^{-1/2}c_2^{-1}\partial_2 F_\kappa(c_1,c_2)}\right) \\
&V_2^{(\infty,\infty)}(c_1,c_2,s) := \\
&c_1^2 + c_2^2 - \mathop{\mathbf{E}}_{(\Lambda,W,G)\sim\mathcal{Q}_\infty}\left(\frac{\Lambda^{-1/2}\,\mathbf{prox}_s\left(\Lambda^{1/2}\Pi_{W^\perp}(G) + \psi^{-1/2}[\partial_1 F_\kappa(c_1,c_2) - c_1c_2^{-1}\partial_2 F_\kappa(c_1,c_2)]\Lambda^{1/2}W\right)}{\psi^{-1/2}c_2^{-1}\partial_2 F_\kappa(c_1,c_2)}\right)^2 \\
&V_3^{(\infty,\infty)}(c_1,c_2,s) := \\
&1 - \mathop{\mathbf{E}}_{(\Lambda,W,G)\sim\mathcal{Q}_\infty}\left|\frac{\Lambda^{-1}\,\mathbf{prox}_s\left(\Lambda^{1/2}G + \psi^{-1/2}[\partial_1 F_\kappa(c_1,c_2) - c_1c_2^{-1}\partial_2 F_\kappa(c_1,c_2)]\Lambda^{1/2}W\right)}{\psi^{-1/2}c_2^{-1}\partial_2 F_\kappa(c_1,c_2)}\right|,
\end{aligned} \tag{A.2}$$

where $F_\kappa(\cdot,\cdot)$ is given by (2.7).

Then from Proposition 3.1, we immediately obtain the following.

Lemma A.1. *Given any (ψ,κ) such that $\psi > \psi^{\downarrow}(\kappa)$, denote $(c_1^\star, c_2^\star, s^\star) \in \mathbb{R}\times\mathbb{R}_{>0}\times\mathbb{R}_{>0}$ to be the unique solution to the system (3.11). Then for every $\epsilon > 0$, there exists $\delta(\epsilon) > 0$ small enough such that if a triplet $(c_1,c_2,s) \in \mathbb{R}\times\mathbb{R}_{>0}\times\mathbb{R}_{>0}$ satisfies*

$$\begin{aligned} |(c_2\vee 1)^{-1} V_1^{(\infty,\infty)}(c_1,c_2,s)| &\le \delta \\ |(c_2\vee 1)^{-2} V_2^{(\infty,\infty)}(c_1,c_2,s)| &\le \delta \\ |(c_2\vee 1)^{-1} V_3^{(\infty,\infty)}(c_1,c_2,s)| &\le \delta, \end{aligned} \tag{A.3}$$

then, (c_1,c_2,s) must be ϵ-close to $(c_1^\star,c_2^\star,s^\star)$,

$$(c_1,c_2,s) \in \mathcal{B}\Big((c_1^\star,c_2^\star,s^\star),\epsilon\Big)\ . \tag{A.4}$$

We next turn to define different empirical versions of (A.2), which will be used later. To this end, recall that (5.8)

$$\hat{F}_\kappa(c_1,c_2) := \Big(\widehat{\mathbf{E}}_n[(\kappa - c_1 Y Z_1 - c_2 Z_2)_+^2]\Big)^{1/2}, \tag{A.5}$$

and define

$$\begin{aligned} &V_1^{(n,p)}(c_1,c_2,s) := \\ &c_1 + \mathop{\mathbf{E}}_{(\Lambda,W,G)\sim\mathcal{Q}_p}\left(\frac{\Lambda^{-1/2}W\cdot\mathbf{prox}_s\Big(\Lambda^{1/2}\Pi_{W^\perp}(G)+\psi^{-1/2}[\partial_1\widehat{F}_\kappa(c_1,c_2)-c_1c_2^{-1}\partial_2\widehat{F}_\kappa(c_1,c_2)]\Lambda^{1/2}W\Big)}{\psi^{-1/2}c_2^{-1}\partial_2\widehat{F}_\kappa(c_1,c_2)}\right) \\ &V_2^{(n,p)}(c_1,c_2,s) := \\ &c_1^2+c_2^2 - \mathop{\mathbf{E}}_{(\Lambda,W,G)\sim\mathcal{Q}_p}\left(\frac{\Lambda^{-1/2}\,\mathbf{prox}_s\Big(\Lambda^{1/2}\Pi_{W^\perp}(G)+\psi^{-1/2}[\partial_1\widehat{F}_\kappa(c_1,c_2)-c_1c_2^{-1}\partial_2\widehat{F}_\kappa(c_1,c_2)]\Lambda^{1/2}W\Big)}{\psi^{-1/2}c_2^{-1}\partial_2\widehat{F}_\kappa(c_1,c_2)}\right)^2 \\ &V_3^{(n,p)}(c_1,c_2,s) := \\ &1 - \mathop{\mathbf{E}}_{(\Lambda,W,G)\sim\mathcal{Q}_p}\left|\frac{\Lambda^{-1}\,\mathbf{prox}_s\Big(\Lambda^{1/2}G+\psi^{-1/2}[\partial_1\widehat{F}_\kappa(c_1,c_2)-c_1c_2^{-1}\partial_2\widehat{F}_\kappa(c_1,c_2)]\Lambda^{1/2}W\Big)}{\psi^{-1/2}c_2^{-1}\partial_2\widehat{F}_\kappa(c_1,c_2)}\right| \end{aligned} \tag{A.6}$$

Finally, define the functions $V_1^{(\infty,p)}(\cdot,\cdot,\cdot), V_2^{(\infty,p)}(\cdot,\cdot,\cdot), V_3^{(\infty,p)}(\cdot,\cdot,\cdot) : \mathbb{R}^3 \to \mathbb{R}$ as follows

$$\begin{aligned} &V_1^{(\infty,p)}(c_1,c_2,s) := \\ &c_1 + \mathop{\mathbf{E}}_{(\Lambda,W,G)\sim\mathcal{Q}_p}\left(\frac{\Lambda^{-1/2}W\cdot\mathbf{prox}_s\Big(\Lambda^{1/2}\Pi_{W^\perp}(G)+\psi^{-1/2}[\partial_1 F_\kappa(c_1,c_2)-c_1c_2^{-1}\partial_2 F_\kappa(c_1,c_2)]\Lambda^{1/2}W\Big)}{\psi^{-1/2}c_2^{-1}\partial_2 F_\kappa(c_1,c_2)}\right) \\ &V_2^{(\infty,p)}(c_1,c_2,s) := \\ &c_1^2+c_2^2 - \mathop{\mathbf{E}}_{(\Lambda,W,G)\sim\mathcal{Q}_p}\left(\frac{\Lambda^{-1/2}\,\mathbf{prox}_s\Big(\Pi_{W^\perp}(G)+\psi^{-1/2}[\partial_1 F_\kappa(c_1,c_2)-c_1c_2^{-1}\partial_2 F_\kappa(c_1,c_2)]\Lambda^{1/2}W\Big)}{\psi^{-1/2}c_2^{-1}\partial_2 F_\kappa(c_1,c_2)}\right)^2 \\ &V_3^{(\infty,p)}(c_1,c_2,s) := \\ &1 - \mathop{\mathbf{E}}_{(\Lambda,W,G)\sim\mathcal{Q}_p}\left|\frac{\Lambda^{-1}\,\mathbf{prox}_s\Big(\Lambda^{1/2}G+\psi^{-1/2}[\partial_1 F_\kappa(c_1,c_2)-c_1c_2^{-1}\partial_2 F_\kappa(c_1,c_2)]\Lambda^{1/2}W\Big)}{\psi^{-1/2}c_2^{-1}\partial_2 F_\kappa(c_1,c_2)}\right|, \end{aligned} \tag{A.7}$$

Observe $V_i^{(\infty,p)}(\cdot,\cdot,\cdot)$ and $V^{(n,p)}(\cdot,\cdot,\cdot)$ only differs in the following sense: $\widehat{F_\kappa}$ is used in place of F_κ.

With the above preparation, we are now in position to establish (5.11). Recall the finite n,p optimization problem

$$\hat{\xi}^{(n,p)}_{\psi,\kappa}(\lambda,w,g) := \min_{\|\theta\|_1\le\sqrt{p}} \psi^{-1/2}\widehat{F}_\kappa\Big(\langle w,\Lambda^{1/2}\theta\rangle, \|\Pi_{w^\perp}(\Lambda^{1/2}\theta)\|_2\Big) + \frac{1}{\sqrt{p}}\Big\langle \Pi_{w^\perp}(g),\Lambda^{1/2}\theta\Big\rangle, \tag{A.8}$$

and the corresponding infinite-dimensional optimization problem given by

$$\begin{aligned}&\tilde{\xi}^{(\infty,\infty)}_{\psi,\kappa}(\Lambda,W,G) :=\\ &\min_{\|h\|_{L_1(\mathcal{Q}_\infty)}\le 1} \psi^{-1/2}F_\kappa\Big(\langle w,\Lambda^{1/2}h\rangle_{L_2(\mathcal{Q}_\infty)}, \|\Pi_{w^\perp}(\Lambda^{1/2}h)\|_{L_2(\mathcal{Q}_\infty)}\Big) + \Big\langle \Pi_{w^\perp}(G),\Lambda^{1/2}h\Big\rangle_{L_2(\mathcal{Q}_\infty)}.\end{aligned} \tag{A.9}$$

Proposition A.1 (Large n,p limit)**.** *Under the assumptions of Theorem 3.1, almost surely,*

$$\lim_{n\to\infty,p(n)/n=\psi} \hat{\xi}^{(n,p)}_{\psi,\kappa}(\lambda,w,g) = \tilde{\xi}^{(\infty,\infty)}_{\psi,\kappa}(\Lambda,W,G), \tag{A.10}$$

where $(\Lambda,W,G)\sim\mathcal{Q}_\infty$.

Proof of Proposition A.1. To begin with, recall the KKT conditions (A.80)-(A.82), together these establish the following fixed point equations

$$V_1^{(\infty,\infty)}(c_1,c_2,s)=0, V_2^{(\infty,\infty)}(c_1,c_2,s)=0, V_3^{(\infty,\infty)}(c_1,c_2,s)=0.$$

We postpone the derivation of the KKT conditions later so as to not interrupt the flow.

Note that the objective function in (A.8) is not convex in θ (due to $\widehat{F}$). Nonetheless, for any θ that minimizes the objective, the KKT conditions still hold as first-order necessary conditions. Thus, by arguments similar to that in the proof of Proposition 3.1, with $\theta/\sqrt{p},\mathcal{Q}_p,\widehat{F_k}$ replacing $h,\mathcal{Q}_\infty,F_k$, we obtain the finite sample versions

$$V_1^{(n,p)}(c_1,c_2,s)=0, V_2^{(n,p)}(c_1,c_2,s)=0, V_3^{(n,p)}(c_1,c_2,s)=0. \tag{A.11}$$

We claim that almost surely, the following uniform convergence result holds, in the region $c_1\in[0,M], c_2>0, s>0$

$$\begin{aligned}\lim_{n\to\infty,p(n)/n=\psi}\sup_{c_1\in[0,M],c_2>0,s>0}(c_2\vee 1)^{-1}|V_1^{(n,p)}(c_1,c_2,s)-V_1^{(\infty,\infty)}(c_1,c_2,s)|=0\\ \lim_{n\to\infty,p(n)/n=\psi}\sup_{c_1\in[0,M],c_2>0,s>0}(c_2\vee 1)^{-2}|V_2^{(n,p)}(c_1,c_2,s)-V_2^{(\infty,\infty)}(c_1,c_2,s)|=0\\ \lim_{n\to\infty,p(n)/n=\psi}\sup_{c_1\in[0,M],c_2>0,s>0}(c_2\vee 1)^{-1}|V_3^{(n,p)}(c_1,c_2,s)-V_3^{(\infty,\infty)}(c_1,c_2,s)|=0\end{aligned} \tag{A.12}$$

In the following, we will prove the above claims.

The first claim in (A.12). By the triangle inequality,

$$|V_1^{(n,p)}(c_1,c_2,s) - V_1^{(\infty,\infty)}(c_1,c_2,s)| \tag{A.13}$$

$$\leq |V_1^{(n,p)}(c_1,c_2,s) - V_1^{(\infty,p)}(c_1,c_2,s)| + |V_1^{(\infty,p)}(c_1,c_2,s) - V_1^{(\infty,\infty)}(c_1,c_2,s)| \ . \tag{A.14}$$

We start with providing a uniform deviation bound in the region $c_1 \in [0,M], c_2 > 0, s > 0$ for

$$(c_2 \vee 1)^{-1} |V_1^{(n,p)}(c_1,c_2,s) - V_1^{(\infty,p)}(c_1,c_2,s)| \ . \tag{A.15}$$

Note here that c_2, s lie in unbounded regions—such a scenario does not arise in the study of the max-L_2-margin, for instance. Define

$$\hat{C}^{\uparrow} := \psi^{-1/2}[\partial_1 \widehat{F}_\kappa(c_1,c_2) - c_1 c_2^{-1} \partial_2 \widehat{F}_\kappa(c_1,c_2)] \tag{A.16}$$

$$\hat{C}^{\downarrow} := \psi^{-1/2} c_2^{-1} \partial_2 \widehat{F}_\kappa(c_1,c_2) \tag{A.17}$$

and similarly $C^{\uparrow}, C^{\downarrow}$ by replacing $\widehat{F}_\kappa$ by F_κ. By the contraction property of the proximal operator,

$$Eqn.\ (A.15) \leq$$

$$(c_2 \vee 1)^{-1} \left\{ \frac{\|\Lambda^{-1/2} G\|_{L_2(\mathcal{Q}_p)} \|\Lambda^{1/2} W\|_{L_2(\mathcal{Q}_p)} + \|W\|^2_{L_2(\mathcal{Q}_p)} |\hat{C}^{\uparrow}|}{|\hat{C}^{\downarrow} C^{\downarrow}|} |\hat{C}^{\downarrow} - C^{\downarrow}| + \frac{\|W\|^2_{L_2(\mathcal{Q}_p)}}{|C^{\downarrow}|} |\hat{C}^{\uparrow} - C^{\uparrow}| \right\} \ . \tag{A.18}$$

As in Lemma 5.1, divide the range of c_2 into the regions $(0,M]$ and (M,∞) respectively. For $c_2 \in (0,M]$, multiply both the denominator and nominator by c_2^2 to obtain

$$Eqn.\ (A.15) \leq \frac{c_2 L + |c_2 \hat{C}^{\uparrow}| L}{|c_2 \hat{C}^{\downarrow}| |c_2 C^{\downarrow}|} |(c_2 \hat{C}^{\downarrow}) - (c_2 C^{\downarrow})| + \frac{L}{|c_2 C^{\downarrow}|} |(c_2 \hat{C}^{\uparrow}) - (c_2 C^{\uparrow})| \tag{A.19}$$

where $L^{1/2}$ is a uniform upper bound on $\|\Lambda^{-1/2} G\|_{L_2(\mathcal{Q}_p)}$, $\|\Lambda^{1/2} W\|_{L_2(\mathcal{Q}_p)}$, $\|W\|_{L_2(\mathcal{Q}_p)}$ for all p. By Lemma 5.1, we know that w.p. at least $1 - n^{-2}$ for all $|c_1| \leq M, 0 < c_2 \leq M, s > 0$

$$|(c_2 \hat{C}^{\downarrow}) - (c_2 C^{\downarrow})| = \psi^{-1/2} |\partial_2 \widehat{F}_\kappa(c_1,c_2) - \partial_2 F_\kappa(c_1,c_2)| \lesssim \frac{\log n}{\sqrt{n}}$$

$$|(c_2 \hat{C}^{\uparrow}) - (c_2 C^{\uparrow})| \leq \psi^{-1/2} c_2 \cdot |\partial_2 \widehat{F}_\kappa(c_1,c_2) - \partial_2 F_\kappa(c_1,c_2)| + \psi^{-1/2} |c_1| \cdot |\partial_1 \widehat{F}_\kappa(c_1,c_2) - \partial_1 F_\kappa(c_1,c_2)| \lesssim \frac{\log n}{\sqrt{n}}$$

which ensures that w.p. at least $1 - n^{-2}$ for all $|c_1| \leq M, 0 < c_2 \leq M, s > 0$,

$$Eqn.\ (A.15) \leq L' \cdot \frac{\log n}{\sqrt{n}} \ , \tag{A.20}$$

and the upper bound is uniform for all p.

For the second region, $c_2 \in (M,\infty)$, we use the following technique as in Lemma 5.1

$$Eqn.\ (A.15) \leq (c_2 \vee 1)^{-1} \left(c_2 \frac{L + |\hat{C}^{\uparrow}| L}{|(c_2 \hat{C}^{\downarrow}| |c_2 C^{\downarrow}|} |(c_2 \hat{C}^{\downarrow}) - (c_2 C^{\downarrow})| + c_2 \frac{L}{|c_2 C^{\downarrow}|} |\hat{C}^{\uparrow} - C^{\uparrow}| \right) \tag{A.21}$$

$$\leq \frac{L + |\hat{C}^{\uparrow}| L}{|(c_2 \hat{C}^{\downarrow}| |c_2 C^{\downarrow}|} |(c_2 \hat{C}^{\downarrow}) - (c_2 C^{\downarrow})| + \frac{L}{|c_2 C^{\downarrow}|} |\hat{C}^{\uparrow} - C^{\uparrow}| \tag{A.22}$$

By Lemma 5.1, we know that w.p. at least $1-n^{-2}$, uniformly for the region $|c_1| \le M, c_2 > M, s > 0$,

$$|(c_2\hat{C}^{\downarrow}) - (c_2 C^{\downarrow})| = \psi^{-1/2}|\partial_2 \widehat{F}_\kappa(c_1,c_2) - \partial_2 F_\kappa(c_1,c_2)| \lesssim \frac{\log n}{\sqrt{n}}$$
$$|\hat{C}^{\uparrow} - C^{\uparrow}| \le \psi^{-1/2}|\partial_2 \widehat{F}_\kappa(c_1,c_2) - \partial_2 F_\kappa(c_1,c_2)| + \psi^{-1/2}|c_1 c_2^{-1}| \cdot |\partial_1 \widehat{F}_\kappa(c_1,c_2) - \partial_1 F_\kappa(c_1,c_2)| \lesssim \frac{\log n}{\sqrt{n}}$$

since $c_1 c_2^{-1}$ is bounded by 1.

Putting things together, we have established that w.p. at least $1-2n^{-2}$,

$$\sup_{|c_1|\le M, c_2>0, s>0} (c_2 \vee 1)^{-1} |V_1^{(n,p)}(c_1,c_2,s) - V_1^{(\infty,p)}(c_1,c_2,s)| \le \frac{\log n}{\sqrt{n}} \ . \tag{A.23}$$

We remark that the above uniform deviation bound over unbounded region is proved due to a key self-normalization property of the function $\partial_i F_\kappa(c_1,c_2), i = 1,2$, as derived in Lemma 5.1.

We now proceed to bound the second term in (A.13)

$$(c_2 \vee 1)^{-1} |V_1^{(\infty,p)}(c_1,c_2,s) - V_1^{(\infty,\infty)}(c_1,c_2,s)| \tag{A.24}$$
$$= \left| \left(\mathop{\mathbf{E}}_{(\Lambda,W,G)\sim\mathcal{Q}_p} - \mathop{\mathbf{E}}_{(\Lambda,W,G)\sim\mathcal{Q}_\infty} \right) (c_2 \vee 1)^{-1} f_{c_1,c_2,s}(\Lambda,W,G) \right|, \tag{A.25}$$

where

$$f_{c_1,c_2,s}(\Lambda,W,G) := \begin{pmatrix} \Lambda^{-1/2} W \cdot \mathbf{prox}_s\left(\Lambda^{1/2}\Pi_{W^\perp}(G) + \psi^{-1/2}[\partial_1 F_\kappa(c_1,c_2) - c_1 c_2^{-1}\partial_2 F_\kappa(c_1,c_2)]\Lambda^{1/2} W\right) \\ \psi^{-1/2} c_2^{-1} \partial_2 F_\kappa(c_1,c_2) \end{pmatrix} \tag{A.26}$$

Since $\mathcal{Q}_p \overset{W_2}{\Longrightarrow} \mathcal{Q}_\infty$, by Theorem 2.7 and Proposition 2.4 in [2], we know that (1) for any function g that grows at most quadratically,

$$\sup_{\Lambda,W,G} \frac{|g(\Lambda,W,G)|}{1 + \|(\Lambda,W,G)\|_2^2} < \infty \ , \tag{A.27}$$
$$\lim_{p\to\infty} \left| \left(\mathop{\mathbf{E}}_{(\Lambda,W,G)\sim\mathcal{Q}_p} - \mathop{\mathbf{E}}_{(\Lambda,W,G)\sim\mathcal{Q}_\infty} \right) g(\Lambda,W,G) \right| = 0 \ , \tag{A.28}$$

and that (2) $\{\mathcal{Q}_p, p \in \mathbb{N}\}$ is 2-uniformly integrable in the following sense: for any $\epsilon > 0$, there exists R_ϵ such that uniformly for p,

$$\sup_{p\in\mathbb{N}} \int_{\mathbb{R}^3 \setminus \mathcal{B}_{R_\epsilon}} \|(\Lambda,W,G)\|^2 d\mathcal{Q}_p < \epsilon \ . \tag{A.29}$$

Here $\mathbb{R}^3 \setminus \mathcal{B}_{R_\epsilon}$ denotes the complement of a ball of radius R_ϵ centered at zero. Note that (1) has not yet established the uniform convergence that we desire. We now prove it using the structural form of $f_{c_1,c_2,s}(\Lambda,W,G)$.

We first verify that $g_{c_1,c_2,s} := (c_2 \vee 1)^{-1} f_{c_1,c_2,s}$ satisfies the quadratic growth condition uniformly for all $|c_1| \le M, c_2 > 0, s > 0$. Observe that

$$|f_{c_1,c_2,s}(\Lambda, W, G)| \le \frac{|W\Pi_{W^\perp}(G)| + |C^\uparrow||W^2|}{|C^\downarrow|} \le \frac{G^2 + W^2 + |C^\uparrow| \cdot W^2}{|C^\downarrow|} \ .$$

Further, for all $|c_1| \le M, 0 \le c_2 \le M, s \ge 0$, uniformly for Λ, W, G (recall that Λ, W has bounded domain)

$$\frac{(c_2 \vee 1)^{-1}|f_{c_1,c_2,s}(\Lambda, W, G)|}{1 + \Lambda^2 + W^2 + G^2} \le \frac{c_2(G^2 + W^2) + |c_2 C^\uparrow| W^2}{|c_2 C^\downarrow|(1 + \Lambda^2 + W^2 + G^2)} < \infty \ , \tag{A.30}$$

since $|c_2 C^\uparrow|$ is bounded above and $|c_2 C^\downarrow| = \psi^{-1/2}|\partial_2 F_\kappa|$ is bounded below. For the other part where $|c_1| \le M, c_2 > M, s \ge 0$, since $|c_1 c_2^{-1}|$ is bounded and, thus, $|C^\uparrow|$ is bounded, hence

$$\frac{(c_2 \vee 1)^{-1}|f_{c_1,c_2,s}(\Lambda, W, G)|}{1 + \Lambda^2 + W^2 + G^2} \le \frac{(G^2 + W^2) + |C^\uparrow| W^2}{|c_2 C^\downarrow|(1 + \Lambda^2 + W^2 + G^2)} < \infty \ . \tag{A.31}$$

Therefore uniformly over $|c_1| \le M, c_2 > 0, s > 0$, with a universal constant K

$$|g_{c_1,c_2,s}(\Lambda, W, G)| = (c_2 \vee 1)^{-1}|f_{c_1,c_2,s}(\Lambda, W, G)| \le K \cdot \|(\Lambda, W, G)\|^2 \ . \tag{A.32}$$

Note that $g_{c_1,c_2,s}(\Lambda, W, G)$ depends on c_1, c_2, s. We now prove the convergence of $\mathbf{E}_{\mathcal{Q}_p}[g_{c_1,c_2,s}]$ to $\mathbf{E}_{\mathcal{Q}_\infty}[g_{c_1,c_2,s}]$ uniformly over c_1, c_2, s. Recall that $\mathcal{Q}_p$ is 2-uniformly integrable, hence for any fixed $\epsilon > 0$, there exists R_ϵ such that (A.29) holds true. Therefore

$$\left|\int g_{c_1,c_2,s} d\mathcal{Q}_p - \int g_{c_1,c_2,s} d\mathcal{Q}_\infty\right| \le \left|\int_{\mathcal{B}_{R_\epsilon}} g_{c_1,c_2,s}(d\mathcal{Q}_p - d\mathcal{Q}_\infty)\right| + \left|\int_{\mathbb{R}^3 \setminus \mathcal{B}_{R_\epsilon}} g_{c_1,c_2,s}(d\mathcal{Q}_p - d\mathcal{Q}_\infty)\right| \tag{A.33}$$

$$\le \left|\int_{\mathcal{B}_{R_\epsilon}} g_{c_1,c_2,s}(d\mathcal{Q}_p - d\mathcal{Q}_\infty)\right| + \left|\int_{\mathbb{R}^3 \setminus \mathcal{B}_{R_\epsilon}} g_{c_1,c_2,s} d\mathcal{Q}_p\right| + \left|\int_{\mathbb{R}^3 \setminus \mathcal{B}_{R_\epsilon}} g_{c_1,c_2,s} d\mathcal{Q}_\infty\right| \ , \tag{A.34}$$

$$\le \left|\int_{\mathcal{B}_{R_\epsilon}} g_{c_1,c_2,s}(d\mathcal{Q}_p - d\mathcal{Q}_\infty)\right| + 2K\epsilon \tag{A.35}$$

where the last step uses the quadratic growth condition of $g_{c_1,c_2,s}$ in (A.32) uniformly over $|c_1| \le M, c_2 > 0, s > 0$, and 2-uniform integrability (A.29), as

$$\left|\int_{\mathbb{R}^3 \setminus \mathcal{B}_{R_\epsilon}} g_{c_1,c_2,s} d\mathcal{Q}_p\right| \le K \int_{\mathbb{R}^3 \setminus \mathcal{B}_{R_\epsilon}} \|(\Lambda, W, G)\|^2 d\mathcal{Q}_p \le K\epsilon \ . \tag{A.36}$$

Inside a bounded region B_{R_ϵ}, it is easy to see that $g_{c_1,c_2,s}(\Lambda, W, G)$ is Lipschitz in (Λ, W, G) with a uniform Lipschitz constant L_{R_ϵ} regardless of the choice of $|c_1| \le M, c_2 > 0, s > 0$. Therefore we have

$$\left|\int_{\mathcal{B}_{R_\epsilon}} g_{c_1,c_2,s}(d\mathcal{Q}_p - d\mathcal{Q}_\infty)\right| \le L_{R_\epsilon} W_1(\mathcal{Q}_p, \mathcal{Q}_\infty) \le L_{R_\epsilon} W_2(\mathcal{Q}_p, \mathcal{Q}_\infty) \ . \tag{A.37}$$

Now we have proved that for

$$\sup_{|c_1|\le M,c_2>0,s>0} |\int g_{c_1,c_2,s} d\mathcal{Q}_p - \int g_{c_1,c_2,s} d\mathcal{Q}_\infty| \le L_{R_\epsilon} W_2(\mathcal{Q}_p,\mathcal{Q}_\infty) + 2K\epsilon \ , \tag{A.38}$$

$$\lim_{p\to\infty} \sup_{|c_1|\le M,c_2>0,s>0} |\int g_{c_1,c_2,s} d\mathcal{Q}_p - \int g_{c_1,c_2,s} d\mathcal{Q}_\infty| \le 2K\epsilon \ . \tag{A.39}$$

By the fact that ϵ can take an arbitrarily small value, we have proved

$$\lim_{n\to\infty,p(n)/n=\psi} \sup_{|c_1|\le M,c_2>0,s>0} (c_2\vee 1)^{-1} |V_1^{(\infty,p)}(c_1,c_2,s) - V_1^{(\infty,\infty)}(c_1,c_2,s)| = 0 \ , \tag{A.40}$$

which handles the second term in (A.13).

We combine with the analysis of (A.15) and by Borel-Cantelli Lemma obtain that, almost surely

$$\lim_{n\to\infty,p(n)/n=\psi} \sup_{|c_1|\le M,c_2>0,s>0} (c_2\vee 1)^{-1} |V_1^{(n,p)}(c_1,c_2,s) - V_1^{(\infty,p)}(c_1,c_2,s)| = 0 \ . \tag{A.41}$$

Thus we have established that uniformly over $|c_1|\le M, c_2>0, s>0$,

$$\lim_{n\to\infty,p(n)/n=\psi} \sup_{|c_1|\le M,c_2>0,s>0} (c_2\vee 1)^{-1} |V_1^{(n,p)}(c_1,c_2,s) - V_1^{(\infty,p)}(c_1,c_2,s)| = 0, \quad a.s. \tag{A.42}$$

The second claim in (A.12). This step follows similarly to the aforementioned analysis, here we only highlight the differences. Once again,

$$|V_2^{(n,p)}(c_1,c_2,s) - V_2^{(\infty,\infty)}(c_1,c_2,s)| \le |V_2^{(n,p)}(c_1,c_2,s) - V_2^{(\infty,p)}(c_1,c_2,s)| + |V_2^{(\infty,p)}(c_1,c_2,s) - V_2^{(\infty,\infty)}(c_1,c_2,s)| \ . \tag{A.43}$$

Now it suffices to provide a uniform deviation bound for $c_1\in[0,M], c_2>0, s>0$

$$(c_2\vee 1)^{-2} |V_2^{(n,p)}(c_1,c_2,s) - V_2^{(\infty,p)}(c_1,c_2,s)| \tag{A.44}$$

$$\le (c_2\vee 1)^{-2} \mathop{\mathbf{E}}_{(\Lambda,W,G)\sim\mathcal{Q}_p} \left\{ \left(\frac{|\Pi_{W^\perp}(G)| + |\hat{C}^\uparrow||W|}{|\hat{C}^\downarrow|} + \frac{|\Pi_{W^\perp}(G)| + |C^\uparrow||W|}{|C^\downarrow|} \right) \right.$$

$$\left. \times \left(\frac{|\Pi_{W^\perp}(G)| + |\hat{C}^\uparrow||W|}{|\hat{C}^\downarrow C^\downarrow|} |\hat{C}^\downarrow - C^\downarrow| + \frac{|W|}{|C^\downarrow|} |\hat{C}^\uparrow - C^\uparrow| \right) \right\}. \tag{A.45}$$

Again we divide the range of c_2 into two parts, $(0,M]$ and (M,∞). For the first part, uniformly over $(c_1,c_2)\in[-M,M]\times(0,M]$, Lemma 5.1 shows that

$$|c_2\hat{C}^\downarrow - c_2 C^\downarrow|, |c_2\hat{C}^\uparrow - c_2 C^\uparrow| \lesssim \frac{\log n}{\sqrt{n}} \ . \tag{A.46}$$

For the second part, uniformly over $(c_1,c_2)\in[-M,M]\times(M,\infty)$, Lemma 5.1 shows that

$$|c_2\hat{C}^\downarrow - c_2 C^\downarrow|, |\hat{C}^\uparrow - C^\uparrow| \lesssim \frac{\log n}{\sqrt{n}} \ . \tag{A.47}$$

In either case, one can show that w.p. at least $1-n^{-2}$,

$$\sup_{|c_1|\le M, c_2>0, s>0} (c_2\vee 1)^{-2}|V_2^{(n,p)}(c_1,c_2,s)-V_2^{(\infty,p)}(c_1,c_2,s)| \le L'\cdot\frac{\log n}{\sqrt{n}} \ . \tag{A.48}$$

For the term,

$$(c_2\vee 1)^{-2}|V_2^{(\infty,p)}(c_1,c_2,s)-V_2^{(\infty,\infty)}(c_1,c_2,s)| \tag{A.49}$$

$$= (c_2\vee 1)^{-2}\left|\left(\mathop{\mathbf{E}}_{(\Lambda,W,G)\sim\mathcal{Q}_p}-\mathop{\mathbf{E}}_{(\Lambda,W,G)\sim\mathcal{Q}_\infty}\right)\tilde{f}_{c_1,c_2,s}(\Lambda,W,G)\right|, \tag{A.50}$$

with

$$\tilde{f}_{c_1,c_2,s}(\Lambda,W,G) := \left(\frac{\Lambda^{-1/2}\,\mathbf{prox}_s\left(\Lambda^{1/2}\Pi_{W^\perp}(G)+\psi^{-1/2}[\partial_1 F_\kappa(c_1,c_2)-c_1c_2^{-1}\partial_2 F_\kappa(c_1,c_2)]\Lambda^{1/2}W\right)}{\psi^{-1/2}c_2^{-1},\partial_2 F_\kappa(c_1,c_2)}\right)^2 \tag{A.51}$$

one can verify that uniformly over $|c_1|\le M, c_2>0, s>0$ and Λ, W, G

$$\frac{(c_2\vee 1)^{-2}|\tilde{f}_{c_1,c_2,s}(\Lambda,W,G)|}{1+\Lambda^2+W^2+G^2} < \infty \ . \tag{A.52}$$

The uniform convergence can be established repeating the argument in (A.33). Therefore,

$$\lim_{n\to\infty,p(n)/n=\psi}\sup_{|c_1|\le M,c_2>0,s>0}(c_2\vee 1)^{-2}|V_2^{(\infty,p)}(c_1,c_2,s)-V_2^{(\infty,\infty)}(c_1,c_2,s)| = 0, \tag{A.53}$$

$$\lim_{n\to\infty,p(n)/n=\psi}\sup_{|c_1|\le M,c_2>0,s>0}(c_2\vee 1)^{-2}|V_2^{(n,p)}(c_1,c_2,s)-V_2^{(\infty,p)}(c_1,c_2,s)| = 0 \ \text{ a.s.} \ . \tag{A.54}$$

The third claim in (A.12)**.** The proof of the following uniform convergence for the term involving V_3 follows the exact same steps as for V_1 and, is therefore, omitted.

We next establish that for any solution $\hat{c}_1,\hat{c}_2,\hat{s}$ that solves the empirical fixed point equation,

$$V_i^{(n,p)}(\hat{c}_1,\hat{c}_2,\hat{s}) = 0$$

one must have that

$$\lim_{n\to\infty,p(n)/n=\psi}\hat{c}_1 = c_1^\star, \qquad \lim_{n\to\infty,p(n)/n=\psi}\hat{c}_2 = c_2^\star, \qquad \lim_{n\to\infty,p(n)/n=\psi}\hat{s} = s^\star \tag{A.55}$$

where $(c_1^\star,c_2^\star,s^\star)$ is the unique solution for the fixed point equation

$$V_i^{(\infty,\infty)}(c_1^\star,c_2^\star,s^\star) = 0 \ .$$

This follows by standard arguments on combining (A.12) and Lemma A.1. For any $\epsilon>0$, there exist $\delta>0$ small enough, that satisfies Eqn. A.3. By the uniform convergence (A.12), for that particular δ, there exist n,p large enough, such that for $(\hat{c}_1,\hat{c}_2,\hat{s})$

$$(1\vee\hat{c}_2)^{-1}|V_1^{(n,p)}(\hat{c}_1,\hat{c}_2,\hat{s})-V_1^{(\infty,\infty)}(\hat{c}_1,\hat{c}_2,\hat{s})| \le \delta \ . \tag{A.56}$$

Recall that $V_1^{(\infty,\infty)}(\hat{c}_1,\hat{c}_2,\hat{s})=0$, which implies

$$(1\vee\hat{c}_2)^{-1}|V_1^{(\infty,\infty)}(\hat{c}_1,\hat{c}_2,\hat{s})|\le\delta\ , \tag{A.57}$$

therefore we know that for all n,p large enough,

$$(\hat{c}_1,\hat{c}_2,\hat{s})\in\mathcal{B}((c_1^\star,c_2^\star,s^\star),\epsilon)\ . \tag{A.58}$$

Note this holds for arbitrary ϵ. Therefore, we have proved Eqn. (A.55).

We remark that this convergence result implies the following: any optimizer $\hat{\theta}$ of the finite n,p optimization problem $\hat{\xi}^{(n,p)}_{\psi,\kappa}(\lambda,w,g)$ must satisfy the necessary condition

$$\|\hat{\theta}\|^2\asymp\|\Lambda^{1/2}\hat{\theta}\|_2^2=\langle w,\Lambda^{1/2}\hat{\theta}\rangle^2+\|\Pi_{w^\perp}\hat{\theta}\|_2^2=\hat{c}_1^2+\hat{c}_2^2\le 2(c_1^\star)^2+2(c_2^\star)^2<4R^2 \tag{A.59}$$

for some absolute constant $R>0$, for sufficiently large n and p. This established property will be useful in the next paragraph.

Given Eqn. (A.55), one can verify by the KKT condition that the optimal value of finite n,p optimization problem $\hat{\xi}^{(n,p)}_{\psi,\kappa}(\lambda,w,g)$ can be expressed in the form

$$\hat{T}(\psi,\kappa;\hat{c}_1,\hat{c}_2,\hat{s}):=\psi^{-1/2}[\widehat{F}_\kappa(\hat{c}_1,\hat{c}_2)-\hat{c}_1\partial_1\widehat{F}_\kappa(\hat{c}_1,\hat{c}_2)-\hat{c}_2\partial_2\widehat{F}_\kappa(\hat{c}_1,\hat{c}_2)]-\hat{s} \tag{A.60}$$

where $\hat{c}_1,\hat{c}_2,\hat{s}$ are solutions to the empirical fixed point equations $V_i^{(n,p)}(\hat{c}_1,\hat{c}_2,\hat{s})=0, i=1,2,3$ (that may not be unique for fixed n,p). Now recall that we have proved for sufficiently large n,p, $\hat{c}_1,\hat{c}_2$ lie in a neighborhood of fixed radius R (does not grow with n,p) around $c_1^\star,c_2^\star$, say denoted by $\mathcal{B}(c_1^\star,R),\mathcal{B}(c_2^\star,R)$. It is easy to show that $\widehat{F}_\kappa$ satisfies the uniform convergence bound

$$\lim_{n\to\infty,p(n)/n=\psi}\ \sup_{c_1,c_2\in\mathcal{B}(c_1^\star,R),\mathcal{B}(c_2^\star,R)}|\widehat{F}_\kappa(c_1,c_2)-F_\kappa(c_1,c_2)|=0\quad a.s. \tag{A.61}$$

By Lemma 5.1, $\partial_1\widehat{F}_\kappa$ and $\partial_2\widehat{F}_\kappa$ all satisfy uniform convergence over $|c_1|\le M, c_2>0$. Therefore

$$\lim_{n\to\infty,p(n)/n=\psi}\ \hat{T}(\psi,\kappa;\hat{c}_1,\hat{c}_2,\hat{s}) \tag{A.62}$$

$$=\lim_{n\to\infty,p(n)/n=\psi}\ \psi^{-1/2}[F_\kappa(\hat{c}_1,\hat{c}_2)-\hat{c}_1\partial_1F_\kappa(\hat{c}_1,\hat{c}_2)-\hat{c}_2\partial_2F_\kappa(\hat{c}_1,\hat{c}_2)]-\hat{s}\quad\text{by uniform convergence} \tag{A.63}$$

$$=\psi^{-1/2}[F_\kappa(c_1^\star,c_2^\star)-c_1^\star\partial_1F_\kappa(c_1^\star,c_2^\star)-c_2^\star\partial_2F_\kappa(c_1^\star,c_2^\star)]-s^\star=T(\psi,\kappa)\ . \tag{A.64}$$

Recall from Corollary A.1 that the RHS equals $\bar{\xi}^{(\infty,\infty)}_{\psi,\kappa}(\Lambda,W,G)$. Therefore, we have shown that the LHS limit exists and is unique. Therefore

$$\begin{aligned}\lim_{n\to\infty,p(n)/n=\psi}\hat{\xi}^{(n,p)}_{\psi,\kappa}(\lambda,w,g)&=\lim_{n\to\infty,p(n)/n=\psi}\hat{T}(\psi,\kappa;\hat{c}_1,\hat{c}_2,\hat{s})\\&=T(\psi,\kappa)=\bar{\xi}^{(\infty,\infty)}_{\psi,\kappa}(\Lambda,W,G)\ .\end{aligned}$$

□

Below, we introduce a key lemma used in the uniform convergence proof in Proposition A.1. This new lemma appears to be new to the literature.

Lemma A.2 (Self-normalization and uniform deviation, Lemma 5.1)**.** *For $i = 1, 2$, we have with probability at least $1 - n^{-2}$,*

$$\sup_{|c_1| \le M, c_2 > 0} |\partial_i \widehat{F}_\kappa(c_1, c_2) - \partial_i F_\kappa(c_1, c_2)| \le C \cdot \frac{\log n}{\sqrt{n}} \ , \tag{A.65}$$

where C is a constant that does not depend on n.

Proof of Lemma 5.1. The proof uses a key self-normalization property of the partial derivatives of F_κ, that ensure good concentration behavior even when c_2 is large. We remark that this structural property makes our uniform convergence result over unbounded region possible in Proposition A.1. Note that

$$\partial_1 \widehat{F}_\kappa(c_1, c_2) = -\frac{\widehat{\mathbf{E}}_n[YZ_1\sigma(\kappa - c_1 YZ_1 - c_2 Z_2)]}{(\widehat{\mathbf{E}}_n[\sigma^2(\kappa - c_1 YZ_1 - c_2 Z_2)])^{1/2}} \ , \tag{A.66}$$

$$\partial_2 \widehat{F}_\kappa(c_1, c_2) = -\frac{\widehat{\mathbf{E}}_n[Z_2\sigma(\kappa - c_1 YZ_1 - c_2 Z_2)]}{(\widehat{\mathbf{E}}_n[\sigma^2(\kappa - c_1 YZ_1 - c_2 Z_2)])^{1/2}} \ , \tag{A.67}$$

where $\sigma(t) := \max(t, 0)$ satisfies the positive homogeneity $\sigma(|c|t) = |c|\sigma(t)$.

We prove the claim by dividing c_2 into two regions, $(0, M]$ and (M, ∞).

In the first region, where $(c_1, c_2) \in [-M, M] \times (0, M]$, it is easy to verify that $R_1(c_1, c_2) := YZ_1\sigma(\kappa - c_1 YZ_1 - c_2 Z_2)$, $R_2(c_1, c_2) := Z_2\sigma(\kappa - c_1 YZ_1 - c_2 Z_2)$ and $R_0(c_1, c_2) := \sigma^2(\kappa - c_1 YZ_1 - c_2 Z_2)$ are all sub-exponential random variables with sub-exponential parameters being at most a constant (depends on M), since $\sigma(\kappa - c_1 YZ_1 - c_2 Z_2), YZ_1, Z_2$ are all sub-Gaussian random variables. Denote the ϵ-covering net as $\mathcal{N}_\epsilon([-M, M] \times (0, M])$, we know that on this bounded region, with probability at least $1 - n^{-2}$,

$$\sup_{(c_1,c_2)\in[-M,M]\times(0,M]} \left|\widehat{\mathbf{E}}_n[R_j(c_1, c_2)] - \mathbf{E}[R_j(c_1, c_2)]\right| \tag{A.68}$$

$$\begin{aligned} &\le \sup_{(c_1',c_2')\in\mathcal{N}_\epsilon} \left|\widehat{\mathbf{E}}_n[R_j(c_1', c_2')] - \mathbf{E}[R_j(c_1', c_2')]\right| + \sup_{(c_1,c_2)\in[-M,M]\times(0,M]} \inf_{(c_1',c_2')\in\mathcal{N}_\epsilon} \left|\widehat{\mathbf{E}}_n[R_j(c_1, c_2)] - \widehat{\mathbf{E}}_n[R_j(c_1', c_2')]\right| \\ &\quad + \sup_{(c_1,c_2)\in[-M,M]\times(0,M]} \inf_{(c_1',c_2')\in\mathcal{N}_\epsilon} \left|\mathbf{E}[R_j(c_1, c_2)] - \mathbf{E}[R_j(c_1', c_2')]\right| \end{aligned} \tag{A.69}$$

$$\lesssim \frac{\log \frac{1}{\epsilon^2}}{\sqrt{n}} + (\log n + 1)\epsilon \lesssim \frac{\log n}{\sqrt{n}}, \ \forall j \in 0, 1, 2 \ . \tag{A.70}$$

The above bound is derived with $\epsilon \asymp 1/\sqrt{n}$. Recall that $\mathbf{E}[R_0(c_1, c_2)] = F_\kappa(c_1, c_2) > 0$. Then for n large enough, the claim follows since

$$\begin{aligned} &|\partial_1 \widehat{F}_\kappa(c_1, c_2) - \partial_1 F_\kappa(c_1, c_2)| \\ &\le \frac{\left|\widehat{\mathbf{E}}_n[R_1(c_1, c_2)] - \mathbf{E}[R_1(c_1, c_2)]\right|}{\sqrt{\mathbf{E}[R_0(c_1, c_2)]}} + \frac{\left|\sqrt{\widehat{\mathbf{E}}_n[R_0(c_1, c_2)]} - \sqrt{\mathbf{E}[R_0(c_1, c_2)]}\right| \cdot \left|\widehat{\mathbf{E}}_n[R_1(c_1, c_2)]\right|}{\sqrt{\mathbf{E}[R_0(c_1, c_2)]\widehat{\mathbf{E}}_n[R_0(c_1, c_2)]}} \lesssim \frac{\log n}{\sqrt{n}} \end{aligned}$$

w.p. at least $1-n^{-2}$ uniformly for all $|c_1| \le M, 0 < c_2 \le M$.

For the second region (unbounded), where $(c_1, c_2) \in [-M, M] \times (M, \infty)$, we use the following self-normalization property of $\partial_i \widehat{F}_\kappa(c_1, c_2)$

$$\partial_1 \widehat{F}_\kappa(c_1, c_2) = -\frac{\widehat{\mathbf{E}}_n[YZ_1\sigma(\kappa c_2^{-1} - c_1 c_2^{-1} YZ_1 - Z_2)]}{(\widehat{\mathbf{E}}_n[\sigma^2(\kappa c_2^{-1} - c_1 c_2^{-1} YZ_1 - Z_2)])^{1/2}} \ , \tag{A.71}$$

$$\partial_2 \widehat{F}_\kappa(c_1, c_2) = -\frac{\widehat{\mathbf{E}}_n[Z_2\sigma(\kappa c_2^{-1} - c_1 c_2^{-1} YZ_1 - Z_2)]}{(\widehat{\mathbf{E}}_n[\sigma^2(\kappa c_2^{-1} - c_1 c_2^{-1} YZ_1 - Z_2)])^{1/2}} \ . \tag{A.72}$$

Now the regions for the parameters of interest are bounded since

$$(c_2^{-1}, c_1 c_2^{-1}) \in [0, 1/M) \times (-1, 1). \tag{A.73}$$

Now define $a = c_2^{-1}, b = c_1 c_2^{-1}$, $\tilde{R}_1(a,b) := YZ_1\sigma(\kappa a - bYZ_1 - Z_2)$, $\tilde{R}_2(a,b) := Z_2\sigma(\kappa a - bYZ_1 - Z_2)$ and $\tilde{R}_0(a,b) := \sigma^2(\kappa a - bYZ_1 - Z_2)$ are all sub-exponential random variables with sub-exponential parameters being at most a constant on the region $(c_1, c_2) \in [-M, M] \times (M, \infty)$. A standard ϵ-covering $\mathcal{N}_\epsilon([0, 1/M) \times (-1, 1))$ on $(a, b) := (c_2^{-1}, c_1 c_2^{-1})$ completes the proof for the region $(c_1, c_2) \in [-M, M] \times (M, \infty)$, since

$$\sup_{(c_1,c_2)\in[-M,M]\times(M,\infty)} \left|\widehat{\mathbf{E}}_n[\tilde{R}_j(c_2^{-1}, c_1 c_2^{-1})] - \mathbf{E}[\tilde{R}_j(c_2^{-1}, c_1 c_2^{-1})]\right| \tag{A.74}$$

$$\le \sup_{(a,b)\in\mathcal{N}_\epsilon} \left|\widehat{\mathbf{E}}_n[\tilde{R}_j(a,b)] - \mathbf{E}[\tilde{R}_j(a,b)]\right| + \sup_{(c_1,c_2)\in[-M,M]\times(M,\infty)} \inf_{(a,b)\in\mathcal{N}_\epsilon} \left|\widehat{\mathbf{E}}_n[\tilde{R}_j(c_2^{-1}, c_1 c_2^{-1})] - \widehat{\mathbf{E}}_n[\tilde{R}_j(a,b)]\right|$$

$$+ \sup_{(c_1,c_2)\in[-M,M]\times(M,\infty)} \inf_{(a,b)\in\mathcal{N}_\epsilon} \left|\mathbf{E}[\tilde{R}_j(c_2^{-1}, c_1 c_2^{-1})] - \mathbf{E}[\tilde{R}_j(a,b)]\right| \tag{A.75}$$

$$\lesssim \frac{\log \frac{1}{\epsilon^2}}{\sqrt{n}} + (\log n + 1)\epsilon \lesssim \frac{\log n}{\sqrt{n}}, \ \forall j \in 0, 1, 2 \ . \tag{A.76}$$

The proof can be completed following standard algebra based on the expression (A.71) and (A.72), since

$$\partial_j \widehat{F}_\kappa(c_1, c_2) = -\frac{\widehat{\mathbf{E}}_n[\tilde{R}_j(c_2^{-1}, c_1 c_2^{-1})]}{\sqrt{\widehat{\mathbf{E}}_n[\tilde{R}_0(c_2^{-1}, c_1 c_2^{-1})]}} \ . \tag{A.77}$$

□

Proof of Theorem 3.2. The proof follows by an adaptation of [67, Section E], utilizing the bounds (A.59).

□

Proof of Corollary 3.3. The proof follows from line by line adaptations of the proof in Section A.2, with the system of equations stated using the new proximal operator in Corollary 3.3. □

A.3 Uniqueness Results

We next present the proof of Proposition 3.1.

Proof of Proposition 3.1. To analyze the equation system (3.11), we will, in fact, begin by examining the objective function in (A.9) as a function of h, that is, define

$$\mathcal{R}_{\psi,\kappa,Q_\infty} = \psi^{-1/2} F_\kappa\left(\langle w, \Lambda^{1/2} h\rangle_{L_2(\mathcal{Q}_\infty)}, \|\Pi_{w^\perp}(\Lambda^{1/2} h)\|_{L_2(\mathcal{Q}_\infty)}\right) + \left\langle \Pi_{w^\perp}(G), \Lambda^{1/2} h\right\rangle_{L_2(\mathcal{Q}_\infty)},$$

and consider the optimization problem

$$\text{minimize} \qquad \mathcal{R}_{\psi,\kappa,Q_\infty}(h) \qquad \text{s.t.} \qquad \|h\|_{L_1(\mathcal{Q}_\infty)} \le 1. \tag{A.78}$$

By arguments similar to that in [67, Section B.3.1], one can show that the function $h \to \mathcal{R}_{\psi,\kappa,Q_\infty}$ is strictly convex and, the minimum of the optimization problem (A.78) is achieved at a unique function $h^\star \in L_2(Q_\infty)$. Then the unique minimizer is determined by the KKT conditions, which in this case can be expressed as

$$\begin{aligned} \Lambda^{1/2}\Pi_{W^\perp}(G) + \psi^{-1/2}\Lambda^{1/2}\left[\partial_1 F_\kappa(c_1,c_2) W + \partial_2 F_\kappa(c_1,c_2)\Pi_{W^\perp}(Z)\right] + s\cdot\partial\|h\|_{L_1(\mathcal{Q}_\infty)} = 0 \ , \\ s(1-\|h\|_{L_1(\mathcal{Q}_\infty)}) = 0 \ , \\ s \ge 0, \|h\|_{L_1(\mathcal{Q}_\infty)} \le 1 \ . \end{aligned} \tag{A.79}$$

Above, Z is given by

$$Z = \begin{cases} \frac{\Pi_{W^\perp}(\Lambda^{1/2}h)}{\|\Pi_{W^\perp}(\Lambda^{1/2}h)\|} & \text{if } \|\Pi_{W^\perp}(\Lambda^{1/2}h)\| > 0 \\ Z' \quad \text{s.t.} \quad \|Z'\| \le 1 & \text{if } \|\Pi_{W^\perp}(\Lambda^{1/2}h)\| = 0 \end{cases}.$$

Now, if $\psi > \psi^{\downarrow}(\kappa)$, and Assumptions 1-3 are satisfied, the conditions B1-B3 in [67, Lemma B.4] are satisfied with $\zeta = \left(\mathop{\mathbf{E}}_{(\Lambda,W)\sim\mu} |\Lambda^{-1/2} W|\right)^{-1}$. Note that this is different from the choice of ζ considered in [67]. With this choice, an adaptation of the arguments in [67, Section B.3.3] with appropriate changes in the constants $M, \Delta, \gamma_+(\Delta), \gamma_-(\Delta), \tilde{\gamma}_+(\Delta)$ and $\tilde{\gamma}_-(\Delta)$ yields that for any minimizer h and the corresponding dual variable s, we have $s > 0$ and $\|\Pi_{W^\perp}(\Lambda^{1/2}h)\| > 0$. Denoting $c_1 := \langle \Lambda^{1/2}h, W\rangle_{L_2(\mathcal{Q})}$ and $c_2 := \|\Pi_{W^\perp}(\Lambda^{1/2}h)\|_{L_2(\mathcal{Q})}$, the KKT conditions can then be rewritten as

$$\begin{aligned} \Pi_{W^\perp}(G) + \psi^{-1/2}\left[\partial_1 F_\kappa(c_1,c_2) W + c_2^{-1}\partial_2 F_\kappa(c_1,c_2)(\Lambda^{1/2}h - c_1 W)\right] + s\cdot\Lambda^{-1/2}\partial\|h\|_{L_1(\mathcal{Q}_\infty)} = 0 \\ \|h\|_{L_1(\mathcal{Q}_\infty)} = 1. \end{aligned} \tag{A.80}$$

From the properties of the proximal mapping operator, the above implies that the unique solution $h^\star$ obeys

$$h^\star = -\frac{\Lambda^{-1}\,\mathbf{prox}_s\left(\Lambda^{1/2}G + \psi^{-1/2}[\partial_1 F_\kappa(c_1,c_2) - c_1 c_2^{-1}\partial_2 F_\kappa(c_1,c_2)]\Lambda^{1/2}W\right)}{\psi^{-1/2}c_2^{-1}\partial_2 F_\kappa(c_1,c_2)}. \tag{A.81}$$

Plugging this in the system

$$c_1 = \langle \Lambda^{1/2}h^\star, W\rangle_{L_2(\mathcal{Q}_\infty)}, \qquad c_1^2 + c_2^2 = \|\Lambda^{1/2}h^\star\|^2_{L_2(\mathcal{Q}_\infty)}, \qquad \|h^\star\|_{L_1(\mathcal{Q}_\infty)} = 1 \tag{A.82}$$

yields the fixed point equations (3.11). Since the solution $h^\star$ is unique, the values $c_1 := \langle \Lambda^{1/2}h^\star, W\rangle_{L_2(\mathcal{Q}_\infty)}$, $c_2 := \|\Pi_{W^\perp}(\Lambda^{1/2}h^\star)\|_{L_2(\mathcal{Q}_\infty)}$ and the value s satisfying (A.80) are also unique and, furthermore, c_2 and s are strictly positive. □

We obtain a key representation for $\tilde{\xi}^{(\infty,\infty)}_{\psi,\kappa}(\Lambda, W, G)$ as a byproduct of the above. On taking inner products with $\Lambda^{1/2}h$ on both sides of the first equation in (A.80) leads to the following.

Corollary A.1. *Under the assumptions of Proposition 3.1, the minimum value of the optimization problem (A.78) is given by*

$$\tilde{\xi}^{(\infty,\infty)}_{\psi,\kappa}(\Lambda, W, G) = \psi^{-1/2}\left[F_\kappa(c_1,c_2) - c_1\partial_1 F_\kappa(c_1,c_2) - c_2\partial_2 F_\kappa(c_1,c_2)\right] - s,$$

where $(c_1, c_2, s) \in \mathbb{R}\times\mathbb{R}_{>0}\times\mathbb{R}_{>0}$ forms the unique solution to (3.11). Hence, the above equals $T(\psi,\kappa)$ defined in (3.1).

Remark A.1. *For the setting of Corollary 3.3, note that $F_\kappa(\cdot,\cdot)$ remains the same as that in the case of the ℓ_1 geometry. Therefore, the arguments in Section A.2 naturally extend to the ℓ_q geometry (as long as $q \le 2$), and those in the current section can also be extended to show uniqueness of the system (3.22) on changing the definition of ζ appropriately. For the settings in Section 3.5, the key function $F_\kappa(\cdot,\cdot)$ changes, however, both (3.25) and (3.27) exhibit similar self-normalization properties that were crucial in the uniform deviation arguments of Section 3.5. Thus, our proofs once again extend naturally to these settings.*

A.4 Proof of Theorem 3.3

Note that the min-ℓ_1-norm interpolated classifier, (1.2), may be expressed as

$$\hat{\theta}_{n,\ell_1} = \arg\min_{\|\theta\|_1 \le \sqrt{p}} \frac{1}{\sqrt{p}} \|(p^{1/2}\kappa_{n,\ell_1}\mathbf{1} - (y\odot X)\theta)_+\|_2 \ . \tag{A.83}$$

The objective function above is the same as $\xi^{(n,p)}_{\psi,\kappa}$ from (5.1), when κ equals $p^{1/2}\kappa_{n,\ell_1}$. For simplicity of notation, we denote this objective function by $\xi^{(n,p)}_{\psi,\tilde{\kappa}}$. To study (A.83), we will utilize the following simple observation [69, Lemma 1]

Lemma A.3. *Suppose $g(\cdot)$ is a convex function and $x^\star = \min_{x\in D} g(x)$ where D is convex. Further, let $\phi(\cdot)$ be such that $g(\cdot) + \epsilon\phi(\cdot)$ remains convex for all $\epsilon \in [-e, e]$, where $e > 0$. Define $\Phi(\epsilon)$ to be the minimal value of $g + \epsilon\phi$ on D. Then $\Phi(\epsilon)$ is concave on $[-e, e]$ and $\phi(x^\star)$ is its subgradient at $\epsilon = 0$.*

This motivates us to define

$$\xi^{(n,p)}_{\psi,\tilde{\kappa},\phi} = \xi^{(n,p)}_{\psi,\tilde{\kappa}} + \epsilon\phi(\theta, \theta_\star), \tag{A.84}$$

where $\phi : \mathbb{R}^p\times\mathbb{R}^p \to \mathbb{R}$ is of the form $\phi(x,y) = \frac{1}{p}\sum_{i=1}^p \phi_0(\sqrt{p}x_i, \sqrt{p}y_i)$.

Using CGMT, and similar techniques as that for Proposition A.1, we can show that

$$\lim_{n\to\infty, p(n)/n=\psi} \xi^{(n,p)}_{\psi,\tilde{\kappa},\phi} \overset{\text{a.s.}}{=} \tilde{\xi}^{(\infty,\infty)}_{\psi,\kappa_\star}(\Lambda, W, G), \tag{A.85}$$

where

$$\begin{aligned}
&\tilde{\xi}^{(\infty,\infty)}_{\psi,\kappa_\star}(\Lambda, W, G)\\
&= \min_{\|h\|_{L_1(\mathcal{Q})}\le 1}\Big[\psi^{-1/2}F_{\kappa_\star(\psi,\rho,\mu)}\left(\langle W, \Lambda^{1/2}h\rangle_{L_2(\mathcal{Q})}, \|\Pi_{W^\perp}(\Lambda^{1/2}h)\|_{L_2(\mathcal{Q})}\right) + \left\langle \Pi_{W^\perp}(G), \Lambda^{1/2}h\right\rangle_{L_2(\mathcal{Q})} + \epsilon\tilde{\phi}(h, h_0)\Big],
\end{aligned} \tag{A.86}$$

where $\tilde{\phi}(\cdot)$ is the analogue of $\phi(\cdot)$ obtained by identifying $\mathcal{L}_2(Q)$ with $\mathbb{R}^p$, with h_0 being the element corresponding to the sequence $\theta_\star/\sqrt{p}$. Denote the corresponding analogue of ϕ_0 to be $\tilde{\phi}_0$. For simplicity of notation, denote $F_{\kappa_\star(\psi,\rho,\mu)}$ to be $F_{\kappa_\star}$.

Lemma A.3 and the definition (A.83) then leads to the following: any function $\phi(\cdot)$ that preserves the convexity of the RHS of (A.84), $\forall \epsilon \in [-\eta, \eta]$ for some $\eta > 0$, obeys

$$\phi(\hat{\theta}_{n,\ell_1}, \theta_\star) \overset{\text{a.s.}}{\to} \frac{\mathrm{d}\tilde{\xi}^{(\infty,\infty)}_{\psi,\kappa_\star}(\Lambda, W, G)}{\mathrm{d}\epsilon}\Big|_{\epsilon=0} \tag{A.87}$$

By standard arguments using KKT conditions (see Appendix A.3), it can be shown that the function that optimizes (A.86) takes the form

$$h^\star_\epsilon = -\frac{\Lambda^{-1}\,\mathbf{prox}_\rho\left(\Lambda^{1/2}G + \psi^{-1/2}[\partial_1 F_{\kappa_\star}(\tilde{c}_1,\tilde{c}_2) - \tilde{c}_1\tilde{c}_2^{-1}\partial_2 F_{\kappa_\star}(\tilde{c}_1,\tilde{c}_2)]\Lambda^{1/2}W\right)}{\psi^{-1/2}\tilde{c}_2^{-1}\partial_2 F_{\kappa_\star}(\tilde{c}_1,\tilde{c}_2)}, \tag{A.88}$$

where

$$\tilde{c}_1 = \langle \Lambda^{1/2}h^\star_\epsilon, W\rangle_{L_2(\mathcal{Q}_\infty)}, \qquad \tilde{c}_1^2 + \tilde{c}_2^2 = \|\Lambda^{1/2}h^\star_\epsilon\|^2_{L_2(\mathcal{Q}_\infty)}, \qquad \|h^\star_\epsilon\|_{L_1(\mathcal{Q}_\infty)} = 1$$

and $\mathbf{prox}_\rho$ is the proximal mapping operator of

$$\rho(x) = s|x| + \epsilon\tilde{\phi}_0(x, \rho\Lambda^{-1/2}W).$$

Calculating the precise value of $\tilde{\xi}^{(\infty,\infty)}_{\psi,\kappa_\star}(\Lambda, W, G)$ by plugging in $h^\star_\epsilon$, (A.88), and differentiating to compute the RHS of (A.87) then leads to

$$\begin{aligned}
&\phi(\hat{\theta}_{n,\ell_1}, \theta_\star) \overset{\text{a.s.}}{\to} \tilde{\phi}(h^\star_\epsilon, h_0)\Big|_{\epsilon=0} \\
&= \mathbb{E}_{(\Lambda,W,G)}\phi_0\left(-\frac{\Lambda^{-1}\,\mathbf{prox}_{s^\star}\left(\Lambda^{1/2}G + \psi^{-1/2}[\partial_1 F_{\kappa_\star}(c_1^\star,c_2^\star) - c_1^\star c_2^{\star-1}\partial_2 F_{\kappa_\star}(c_1^\star,c_2^\star)]\Lambda^{1/2}W\right)}{\psi^{-1/2}c_2^{\star-1}\partial_2 F_{\kappa_\star}(c_1^\star,c_2^\star)}, \rho\Lambda^{-1/2}W\right).
\end{aligned}$$

Above, $(c_1^\star, c_2^\star, s^\star)$ forms the unique solution to (3.11) with $\kappa = \kappa_\star(\psi,\rho,\mu)$.

A.5 Optimization Results

Proof of Proposition 5.1. We will show the convergence of *Boosting Algorithm* as a special instantiation of the *Mirror Descent* proof. We will establish the result for two scenarios: (1) AdaBoost, with $X_{ij} \in \{\pm 1\}$, and (2) *Boosting Algorithm* in Eqn. (2.9), with bounded continuous $|X_{ij}| \le M$ and a shrinkage on the learning rate (the specifics will be made clear in the proof below).

We will need some background before stating the mirror descent proof. For $x \in \mathbb{R}^n$, define the entropy

$$R(x) = \sum_{i=1}^n x[i]\log(x[i]) + \mathbb{I}_{\Delta_n}(x)\ . \tag{A.89}$$

Here $\mathbb{I}_{\Delta_n}$ is the indicator function on the probability simplex Δ_n. The Fenchel conjugate of R, denoted by $R^\star$, reads,

$$R^\star(x) = \log\left(\sum_{i=1}^{n} \exp(x[i])\right) \ . \tag{A.90}$$

One can verify that R is 1-strongly convex w.r.t. the ℓ_1 norm, and that $R^\star$ is 1-strongly smooth w.r.t. the L_∞ norm.

First, let us recall the dual formulation of ℓ_1-margin, and the von Neumann's minimax theorem

$$\kappa_{n,\ell_1} = \max_{\|\theta\|_1 \le 1} \min_{i\in[n]} e_i^\top Z\theta = \min_{\eta\in\Delta_n} \max_{\|\theta\|_1\le 1} \eta^\top Z\theta = \min_{\eta\in\Delta_n} \|Z^\top \eta\|_\infty \ . \tag{A.91}$$

Therefore, for any $\eta \in \Delta_n$, $\kappa_{n,\ell_1} \le \|Z^\top\eta\|_\infty$.

It is easy to verify that the (1) AdaBoost algorithm defined above is equivalent to the following mirror descent algorithm:

- ℓ_1-margin $\gamma_t := \max_{j\in[p]} |\eta_t^\top Z e_j| = \|Z^\top \eta_t\|_\infty \ge \kappa_{n,\ell_1}$;
- Learning rate is $\alpha_t = \frac{1}{2}\log\frac{1+\gamma_t}{1-\gamma_t}$ since

$$\min_{v\in\{\pm e_j\}_{j\in[p]}} \sum_{i\in[n]} \eta_t[i]\cdot \mathbb{I}_{y_i x_i^\top v\le 0} = \min_{v\in\{\pm e_j\}_{j\in[p]}} \sum_{i\in[n]} \eta_t[i]\cdot \mathbb{I}_{-y_i x_i^\top v\ge 0} \tag{A.92}$$

$$= \frac{1}{2}(-\max_{j\in[p]} |\eta_t^\top Z e_j| + 1) \ ; \tag{A.93}$$

- Updates on $\eta_t \in \Delta_n$ (mirror descent) reduce to

$$\nabla R(\eta_t) = -Z\theta_t \ \text{ (map to mirror space)}, \tag{A.94}$$

$$Z\theta_{t+1} = Z\theta_t + \alpha_t Z v_{t+1} \ \text{ (descent step)}, \tag{A.95}$$

$$\nabla R^\star(-Z\theta_{t+1}) = \eta_{t+1} \ \text{ (inverse map)}. \tag{A.96}$$

Now we are ready to prove the final statement. Due to the fact that $R^\star$ is strongly smooth w.r.t. the L_∞ norm

$$\begin{aligned}
&R^\star(-Z\theta_{t+1}) - R^\star(-Z\theta_t) \\
&\le \langle -\alpha_t Z v_{t+1}, \nabla R^\star(-Z\theta_t)\rangle + \frac{1}{2}\|\alpha_t Z v_{t+1}\|_\infty^2 \\
&\le -\alpha_t \langle Z v_{t+1}, \eta_t\rangle + \frac{1}{2}\alpha_t^2 \|Z v_{t+1}\|_\infty^2 \\
&= -\alpha_t \|Z^\top \eta_t\|_\infty + \frac{1}{2}\alpha_t^2 \quad \text{(here we use the fact that } |Z_{ij}|\le 1) \\
&= -\alpha_t\gamma_t + \frac{1}{2}\alpha_t^2 \le -\frac{\gamma_t^2}{2}(1+o(\gamma_t)) \ .
\end{aligned}$$

The above derives the reduction in $R^\star$ for each step.

For the (2) *Boosting Algorithm* in Eqn. (2.9), with $|X_{ij}| \le M$, define a shrinkage on the learning rate $\alpha_t(\beta)$ with a constant factor $\beta > 0$,

$$\alpha_t(\beta) = \beta \cdot \eta_t^\top Z v_{t+1} \ . \tag{A.97}$$

A good choice of β will be clear in a second. Then

$$\begin{aligned}
&R^\star(-Z\theta_{t+1}) - R^\star(-Z\theta_t) \\
&= -\alpha_t(\beta)\|Z^\top \eta_t\|_\infty + \frac{1}{2}\alpha_t^2(\beta)\|Z v_{t+1}\|_\infty^2 \quad \text{(here we use the fact that } |Z_{ij}| \le M) \\
&= -\beta\gamma_t^2 + \frac{M^2}{2}\beta^2\gamma_t^2 = -\frac{\gamma_t^2}{2M^2}
\end{aligned}$$

where the last step uses the choice of $\beta = 1/M^2$.

Now telescoping with the terms $R^\star(-Z\theta_{t+1}) - R^\star(-Z\theta_t)$, we have

$$R^\star(-Z\theta_T) - R^\star(-Z\theta_0) \le -\frac{\sum_{t=0}^{T-1}\gamma_t^2}{2M^2} \le -T\frac{\kappa_{n,\ell_1}^2}{2M^2} \quad \text{(recall } \gamma_t \ge \kappa_{n,\ell_1}) \tag{A.98}$$

$$\sum_{i\in[n]} \mathbb{I}_{-y_i x_i^\top \theta_T > 0} \le \sum_{i\in[n]} \exp(-y_i x_i^\top \theta_T) = \exp(R^\star(-Z\theta_T)) \le ne \cdot \exp(-T\frac{\kappa_{n,\ell_1}^2}{2M^2}) \ . \tag{A.99}$$

The proof is now complete.

□

Proof of Corollary 5.1. The proof follows from Proposition 5.1 and a re-scaling technique in [97]'s asymptotic analysis. Here instead, we spell out a non-asymptotic result. For any $\kappa > 0$

$$\sum_{i\in[n]} \mathbb{I}_{\frac{y_i x_i^\top \theta_t}{\|\theta_t\|_1} \le \kappa} \le \sum_{i\in[n]} \exp(\kappa\|\theta_t\|_1 - y_i x_i^T \theta_t) \ , \tag{A.100}$$

$$\le \exp(\kappa\|\theta_t\|_1)\exp\Big(R^\star(-Z\theta_t)\Big) \ , \tag{A.101}$$

with $R^\star$ defined in (A.90). Due to the proof in Proposition 5.1, we know

$$R^\star(-Z\theta_T) \le R^\star(-Z\theta_0) - \sum_{t=0}^{T-1}\left(\beta\gamma_t^2 - \frac{\beta^2\gamma_t^2}{2}M^2\right) \tag{A.102}$$

$$\le \log(ne) - \sum_{t=0}^{T-1}\beta\gamma_t\left[\gamma_t - \frac{\beta}{2}\gamma_t M^2\right] \ . \tag{A.103}$$

In addition, due to the coordinate update of θ_t, we know

$$\|\theta_T\|_1 \le \sum_{t=0}^{T-1}\|\alpha_t v_{t+1}\|_1 \le \sum_{t=0}^{T-1}\beta\gamma_t \ . \tag{A.104}$$

Therefore

$$\sum_{i\in[n]} \mathbb{I}_{\frac{y_i x_i^\top \theta_t}{\|\theta_t\|_1} \le \kappa} \le ne \cdot \exp\left\{-\sum_{t=0}^{T-1}\beta\gamma_t\left[\gamma_t - \frac{\beta}{2}\gamma_t M^2 - \kappa\right]\right\} \ . \tag{A.105}$$

Recall that $\gamma_t \geq \kappa_{n,\ell_1}$ for all t, we know that

$$\sum_{i\in[n]} \mathbb{I}_{\frac{y_i x_i^\top \theta_t}{\|\theta_t\|_1}\leq\kappa} \leq ne\cdot\exp\left(-T\beta\kappa_{n,\ell_1}\left[\kappa_{n,\ell_1}(1-\frac{\beta M^2}{2})-\kappa\right]\right) \ . \tag{A.106}$$

With the choice of

$$\beta = \frac{1-\kappa/\kappa_{n,\ell_1}}{M^2}, \quad \text{and} \tag{A.107}$$

$$T \geq \log(1.01ne)\cdot\frac{2M^2\kappa_{n,\ell_1}^{-2}}{(1-\kappa/\kappa_{n,\ell_1})^2}, \tag{A.108}$$

we know that

$$\sum_{i\in[n]} \mathbb{I}_{\frac{y_i x_i^\top \theta_t}{\|\theta_t\|_1}\leq\kappa} \leq \frac{1}{1.01} < 1 \ . \tag{A.109}$$

which implies that $\min_{i\in[n]} \frac{y_i x_i^\top \theta_T}{\|\theta_T\|_1} > \kappa$. Therefore for any $\epsilon < 1$, plug in $\kappa = \kappa_{n,\ell_1}\cdot(1-\epsilon)$

$$T \geq \log(1.01ne)\cdot\frac{2M^2\kappa_{n,\ell_1}^{-2}}{\epsilon^2} \ , \tag{A.110}$$

we must have that

$$\min_{i\in[n]} \frac{y_i x_i^\top \theta_T}{\|\theta_T\|_1} > \kappa_{n,\ell_1}\cdot(1-\epsilon) \ . \tag{A.111}$$

□

Proof of Corollary 3.2. The proof follows by modifying some steps of our proof in the $q = 1$ case. Recall the notations in (A.94),

$$\begin{aligned}
&R^\star(-Z\theta_{t+1}) - R^\star(-Z\theta_t) \\
&\leq \langle -\alpha_t Z v_{t+1}, \nabla R^\star(-Z\theta_t)\rangle + \frac{1}{2}\|\alpha_t Z v_{t+1}\|_\infty^2 \\
&\leq -\alpha_t\langle Z v_{t+1}, \eta_t\rangle + \frac{1}{2}\alpha_t^2\|Z v_{t+1}\|_\infty^2 \\
&= -\alpha_t\|Z^\top\eta_t\|_{q_\star} + \frac{1}{2}\alpha_t^2\max_{i\in[n]}|\langle Z_{i\cdot}, v_{t+1}\rangle|^2 \quad \text{(here we use the fact that } |Z_{ij}|\leq M) \\
&\leq -\alpha_t\gamma_t + \frac{M^2 p^{\frac{2}{q_\star}}}{2}\alpha_t^2 = -\beta\gamma_t^2 + \frac{M^2 p^{\frac{2}{q_\star}}}{2}\beta^2\gamma_t^2
\end{aligned}$$

with $\gamma_t = \|Z^\top\eta_t\|_{q_\star}$.

Observe that

$$\|\theta_T\|_q \leq \sum_{t=0}^{T-1}\|\alpha_t v_{t+1}\|_q \leq \sum_{t=0}^{T-1}\beta\gamma_t \ . \tag{A.112}$$

Plug in the above to the argument in (A.105), we have

$$\sum_{i\in[n]} \mathbb{I}_{\frac{y_i x_i^\top \theta_t}{\|\theta_t\|_q}\le\kappa} \le ne\cdot\exp\left\{-\sum_{t=0}^{T-1}\beta\gamma_t\left[\gamma_t - \frac{\beta}{2}\gamma_t M^2 p^{\frac{2}{q_\star}} - \kappa\right]\right\} \tag{A.113}$$

$$\le ne\cdot\exp\left\{-\sum_{t=0}^{T-1}\beta\gamma_t\left[\gamma_t(1 - \frac{\beta}{2}M^2 p^{\frac{2}{q_\star}}) - \kappa\right]\right\} \tag{A.114}$$

$$\le ne\cdot\exp\left(-T\beta\kappa_{n,\ell_q}^2\left[(1 - \frac{\beta}{2}M^2 p^{\frac{2}{q_\star}}) - \frac{\kappa}{\kappa_{n,\ell_q}}\right]\right) \ . \tag{A.115}$$

where the last step uses the Sion's Minimax Theorem,

$$\gamma_t = \|Z^\top \eta_t\|_{q_\star} \ge \min_{\eta\in\Delta}\max_{\|\theta\|_q\le 1} \eta^\top Z\theta = \max_{\|\theta\|_q\le 1}\min_{i\in[n]} y_i x_i^\top\theta = \kappa_{n,\ell_q} \ . \tag{A.116}$$

The proof is complete if we plug in

$$\beta = \frac{1-\kappa/\kappa_{n,\ell_q}}{p^{\frac{2}{q_\star}}M^2}. \tag{A.117}$$

□

For completeness, we show that the min-ℓ_1-norm interpolation, is equivalent to the max-ℓ_1-margin formulation. We use this fact several places in the main text.

Proposition A.2. *The following two formulations are equivalent*

$$\textit{Formulation I:}\quad I^\star := \max\left\{\kappa \mid \exists\theta,\ \|\theta\|_1\le 1,\ \ s.t.\ \ \forall i\le n,\ y_i x_i^\top\theta\ge\kappa\right\} \tag{A.118}$$

$$\textit{Formulation II:}\quad II^\star := \min\ \|\theta\|_1,\ \ s.t.\ \ \forall i\le n,\ y_i x_i^\top\theta\ge 1 \tag{A.119}$$

and that

$$I^\star = 1/II^\star \ .$$

Proof. Suppose that $\theta_\star$ solves II, then take $\theta = \theta_\star/II^\star$ satisfy $\|\theta\| = 1$, then

$$I^\star \ge 1/II^\star \ .$$

Suppose that $I^\star$ is the optimal solution for I, then there exist a $\theta, \|\theta\|\le 1$ such that $y_i x_i^\top(\theta/I^\star)\ge 1$, then

$$II^\star \le \|\theta/I^\star\|_1 \le 1/I^\star \ .$$

□