

# A Precise High-Dimensional Asymptotic Theory for Boosting and Minimum- $\ell_1$ -Norm Interpolated Classifiers

Tengyuan Liang <sup>\*1</sup> and Pragya Sur <sup>†2</sup>

<sup>1</sup>University of Chicago

<sup>2</sup>Harvard University

## Abstract

This paper establishes a precise high-dimensional asymptotic theory for boosting on separable data, taking statistical and computational perspectives. We consider a high-dimensional setting where the number of features (weak learners)  $p$  scales with the sample size  $n$ , in an overparametrized regime. Under a class of statistical models, we provide an exact analysis of the generalization error of boosting when the algorithm interpolates the training data and maximizes the empirical  $\ell_1$ -margin. Further, we explicitly pin down the relation between the boosting test error and the optimal Bayes error, as well as the proportion of active features at interpolation (with zero initialization). In turn, these precise characterizations answer certain questions raised in [16, 89] surrounding boosting, under assumed data generating processes. At the heart of our theory lies an in-depth study of the maximum- $\ell_1$ -margin, which can be accurately described by a new system of non-linear equations; to analyze this margin, we rely on Gaussian comparison techniques and develop a novel uniform deviation argument. Our statistical and computational arguments can handle (1) any finite-rank spiked covariance model for the feature distribution and (2) variants of boosting corresponding to general  $\ell_q$ -geometry,  $q \in [1, 2]$ . As a final component, via the Lindeberg principle, we establish a universality result showcasing that the scaled  $\ell_1$ -margin (asymptotically) remains the same, whether the covariates used for boosting arise from a non-linear random feature model or an appropriately linearized model with matching moments.

## 1 Introduction

Modern machine learning methods are regularly used for classification tasks. Typically, these algorithms are complex, and often produce solutions with zero training error, even for random labels. Prominent examples include ensemble learning, neural networks, and kernel machines.

---

<sup>\*</sup>[tengyuan.liang@chicagobooth.edu](mailto:tengyuan.liang@chicagobooth.edu). T. Liang acknowledges support from the NSF Career award (DMS-2042473), the George C. Tiao faculty fellowship and the William S. Fishman faculty research fund at the University of Chicago Booth School of Business. T. Liang wishes to thank Yoav Freund, Bin Yu, Misha Belkin, as well as participants in the Learning Theory seminar at Google Research and NSF-Simons Collaboration on Mathematics of Deep Learning for constructive feedback that greatly improved the paper.

<sup>†</sup>[pragya@fas.harvard.edu](mailto:pragya@fas.harvard.edu). P. Sur was partially supported by the Center for Research on Computation and Society, Harvard John A. Paulson School of Engineering and Applied Sciences and by NSF DMS-2113426. P. Sur wishes to thank the organizers and participants of the Young Data Science Researcher Seminar, ETH Zurich, for constructive feedback.

However, among the many solutions that interpolate the training data, not all exhibit superior generalization. Empirically, it has been commonly observed that practical algorithms—running even on large overparametrized models—favor minimal ways of interpolating the training data, which has been conjectured to be crucial for good generalization. Different problem formulations and optimization algorithms favor distinct notions of minimalism, typically measured by specific norms of the classifier. This paper focuses on the celebrated boosting/AdaBoost algorithm in this minimum-norm interpolation regime, where we conduct a precise analysis of its statistical and computational properties under specific data-generating mechanisms.

Ensemble learning algorithms, recognized as powerful toolkits at the disposal of a data scientist, have found widespread usage across domains. Boosting is arguably one of the most powerful ensemble learning algorithms that combines weak learners using intelligent schemes and exhibits remarkable generalization performance. The groundbreaking AdaBoost paper, Freund and Schapire [43], is widely regarded as the milestone in the boosting literature, which can be traced back even earlier [88, 42]. 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 [89], which led to decades of research activities devoted to a better understanding of this algorithm and its variants.

The seminal papers [15, 36, 78] observed that AdaBoost achieves zero error on the training data within a few iterations, whereas the generalization error continues to decrease well beyond this interpolation timepoint. Recently, similar phenomena and puzzles resurfaced in the context of neural networks [106], and motivated the study of interpolation and implicit regularization [9, 8, 66, 53, 7, 68]. 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 [89] 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. [89] observed that over the course of iterations, AdaBoost creates classifiers such that the fraction of training examples with a large margin increases, and the empirical margin distribution stabilizes to a limiting one rapidly. In particular, given any margin level  $\kappa > 0$ , they discovered upper bounds on the prediction error that reveal interesting tradeoffs between two terms—(i) the fraction of training examples with margin below  $\kappa$ , and (ii) the term  $\kappa^{-1}C(\mathcal{H})/\sqrt{n}$  that involves the complexity of the class  $C(\mathcal{H})$  and the sample size  $n$  scaled by  $\kappa$ . A large empirical margin distribution was then conjectured to be a key factor behind the superior generalization performance of certain classifiers. These upper bounds provided extremely useful insights, nonetheless, [89] 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 [16] subsequently contended these empirical margin distribution based explanations, using extensive simulations, and proposed to bound the generalization error using the *minimum value of the margin* over the training set. Later, Koltchinskii and Panchenko [62] improved the earlier bounds from [89]. Despite significant progress in this direction, since these results involved upper bounds, the qualitative question regarding key quantities that precisely determine the generalization behavior of AdaBoost remained unanswered.

**Consistency and early stopping.** In conjunction with the generalization error, statisticians and

learning theorists deeply care about the consistency of AdaBoost, and in particular, about the precise relationship between the test error and the optimal Bayes error. The problem of consistency was posed by Breiman [17], who studied convergence properties of the algorithm in the population case. The seminal papers Jiang [59], Lugosi and Vayatis [70], Zhang [107], Koltchinskii and Besnozova [61] considered different function classes and variants of boosting, and furthered this direction of research. [59] 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 [46, 51, 81, 74]—AdaBoost may overfit, particularly in complex model classes and high noise settings, when left to run for an arbitrary large number of steps. On the one hand, these naturally inspired subsequent work on appropriate regularization strategies for “early stopping” as in Zhang and Yu [108], Bartlett and Traskin [6]. On the other hand, as the model classes become complex and overparametrized, the test error of boosting algorithms may deviate from the optimal Bayes error. Despite an extensive bulk of work, a precise characterization of the test error and its relation to the Bayes error for the overparametrized case is still missing in the current literature.

**Connections with min- $\ell_1$ -norm interpolation (and implications).** In a venture to understand the path of boosting iterates better, Rosset, Zhu and Hastie [83], Zhang and Yu [108] established that for linearly separable data, AdaBoost with infinitesimal step size converges to the minimum- $\ell_1$ -norm interpolated classifier (Equation (1.2)) when left to run forever. This interpolant is crucially related to the maximum  $\ell_1$ -margin on the data,  $\kappa_{n,\ell_1}$  (Equation (1.3)). In fact, expressed differently, these results establish that the number of optimization steps necessary for AdaBoost to reach zero training error can be upper bounded by  $O(\kappa_{n,\ell_1}^{-2})$ . Together with the earlier results Breiman [16], this leads to a plausible conjecture that the max- $\ell_1$ -margin is a crucial quantity that determines both generalization and optimization behaviors of boosting algorithms. (See also [98], for methods to shrink step sizes so that AdaBoost produces approximate maximum margin classifiers.) Thus, understanding the precise value of this margin, and the iteration time necessary for convergence to the min- $\ell_1$ -norm interpolant (on separable data) is crucial for settling such a conjecture. Furthermore, refined analyses of such quantities for various overparametrized models is expected to shed light on the effects of overparametrization on optimization, an understanding that has so far eluded the literature.

Rosset et al. [83] further discussed that the aforementioned convergence to min- $\ell_1$ -norm interpolated classifiers indicates the following: boosting potentially converges (in direction) to a sparse classifier. It would then be of interest to understand properties of the limiting solution better, for example, the analyst may wish to understand the number of weak learners deemed important by the boosting solution. This is particularly crucial in today’s context where producing interpretable classifiers in high-stakes decision making has critical social consequences [69, 28, 84, 60, 105]. Boosting has subsequently witnessed widespread development, and varying perspectives have emerged through several seminal works e.g. [46, 47, 21, 18, 85, 41]; see Section 4 for further discussions.

**This paper.** Prior literature suggested that the min- $\ell_1$ -norm interpolated classifier and the max- $\ell_1$ -margin may form central characters behind boosting algorithms on linearly separable data. However, a thorough understanding of their exact relations with the boosting solution, whether these are key quantities, and how these objects behave, have so far been lacking. When there is label noise in  $y$ , conditional on the features  $x$ , linear separability only happens in an overparametrized regime where the number of features  $p$  grows with the sample size  $n$ ; to see this, note that a fixed

$p$ -dimensional linear model class, cannot shatter  $n$ -points with all possible signs when  $n$  grows.

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

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

$$p/n \rightarrow \psi > 0. \quad (1.1)$$

This is a natural high-dimensional setting for analyzing separable data [24, 76], as argued above; this regime has also been investigated for regression problems and other contexts (see for instance, [34, 38, 33, 104, 37, 96, 97, 39], 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  $\psi$  lying above a threshold that depends on the underlying signal strength of the problem [24, 31, 76]; see Section 2 for further details. Define the *min- $\ell_1$ -norm interpolated classifier* to be

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

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

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

whenever  $\kappa_{n,\ell_1}$  is positive. We first study a stylized model where each row of the design matrix follows a Gaussian distribution with a diagonal covariance, the response is binary, and the distribution of the response conditional on the covariates is given by a generalized linear model as in (2.1) (see Section 2 for further details). Later, Section 3.5 presents extensions to showcase that the precise asymptotic theory carries over to spiked covariance models and random feature models. Therefore, we think of the stylized diagonal Gaussian model as capturing the essence of the mathematical derivations without overwhelming the readers, but certainly not the sole situation where precise asymptotics can be derived. In the aforementioned setting, this paper provides the following contributions to the statistical and computational understanding of boosting:

- (i). We characterize precisely the value of the max- $\ell_1$ -margin (Theorem 3.1) in the high-dimensional regime (1.1), answering a question raised in [16]. Informally, we show that  $\sqrt{p}\kappa_{n,\ell_1}$  converges almost surely to a constant  $\kappa_\star$  that depends on  $\psi$  and other problem parameters, such as the signal-to-noise ratio in the data generating model. Theorem 3.1 explicitly pins down the

limiting constant  $\kappa_\star$ ; in fact, this can be entirely described by the fixed points of a complicated yet easy to solve non-linear system of equations that we will introduce in (3.9). This limiting characterization will prove crucial for understanding the properties of boosting on (asymptotically) separable data.

- (ii). In parallel, we establish precise formulae for the generalization error of the min- $\ell_1$ -norm interpolant  $\hat{\theta}_{n,\ell_1}$  (Theorem 3.2), once again in the regime (1.1). The formula illuminates that the generalization error is completely governed by the dimensionality parameter  $\psi$  and the limit  $\kappa_\star$  characterized in the preceding step. The consequences of this result for boosting will be discussed soon; notably, the min- $\ell_1$ -norm interpolant has been conjectured to be crucial in other contexts (see Section 4), and therefore, we expect Theorem 3.2 to be of wider importance beyond boosting.
- (iii). Turning to boosting, we provide an exact characterization of a threshold  $T$  such that for all iterations  $t \geq T$ , the boosting iterates (with a properly scaled step size) stay arbitrarily close to  $\hat{\theta}_{n,\ell_1}$ , in the large  $n, p$  limit (1.1) (Theorem 3.3). This characterization builds upon existing works on margin maximization that provide a  $1/\sqrt{t}$  rate [98, 40], and uses the well-known rescaling technique, shrinkage technique and mirror descent connections of boosting (see [108, 90, 40, 52, 57, 27] for a non-exhaustive set of related works). However, together with Theorems 3.1-3.2, this result provides an exact characterization of the generalization error of boosting, and improves upon the existing upper bounds by a margin [89, 62], in our setting. Crucially, this formula involves  $\kappa_\star$  (through an implicit non-linear function), and therefore, our results imply that, at least under the aforementioned data-generation scheme, the max- $\ell_1$ -margin drives the generalization performance of boosting. Furthermore, the formula encodes a concrete recipe for comparing the test error of boosting with the Bayes error in high dimensions.
- (iv). 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- $\ell_1$ -margin  $\kappa_\star$ . Utilizing this, we demonstrate two curious phenomena regarding overparametrization, both not known earlier for boosting. (1) Keeping other problem parameters fixed,  $T$  decreases with an increase in  $\psi$ , suggesting that *overparametrization helps in optimization*. (2) We establish bounds on the fraction of activated coordinates in the boosting solution (with zero initialization) when it first interpolates the training data.
- (v). Finally, we introduce a new class of boosting algorithms that converge to the max- $\ell_q$ -margin direction (Section 3.4) for  $q > 1$ . [83] discussed the importance of studying such notions of margins, since it is unclear which geometry induces a better solution (see also [52]). Here, we construct such algorithms and provide precise analyses of their generalization (for the case  $q \in [1, 2]$ ) and optimization properties (for all  $q > 1$ ) in a spirit similar to that for boosting done above.

On the theoretical end, our analyses for the above contributions build upon classical results in Gaussian comparison inequalities [49, 50] that have been strengthened relatively recently [92, 99, 102], leading to the *Convex Gaussian Min-Max Theorem* (CGMT) (see Section 4 for a discussion). The topic of max- $\ell_2$ -margin has received considerable attention, dating back to [48, 91], and has more recently been analyzed in [76, 31]. Our proofs begin from these existing



theory surrounding the max- $\ell_2$ -margin, particularly [76, 31], however, the  $\ell_2$  (coordinate invariant) and  $\ell_q$  ( $q \neq 2$ , coordinate specific) geometries differ significantly. Therefore, considerable theoretical work is necessary to obtain the precise characterizations outlined above; our key contributions in this regard are highlighted in Section 5. Specifically, we introduce a novel uniform deviation argument, which later (Section 3.5) allows us to extend our results to settings with non-diagonal covariance between features.

The aforementioned contributions rely on a specific data-generating scheme that, to a curious reader, might appear stylized. However, the qualitative message remains the same in several settings beyond this specific scheme. Section 3.5 explores this in further detail. In particular, we establish similar characterization for the max- $\ell_1$ -margin and the min- $\ell_1$ -norm interpolant for a class of models where the feature covariance is a finite-rank perturbation of a diagonal (see Section 3.5.1 and Appendix B.1, where we call it the spiked covariance model). Our result can in turn be utilized to establish boosting properties analogous to point (iii) above, for these other data generation schemes. We remark that the simplest model in this class—the rank-one perturbation model—corresponds to the standard Gaussian mixture model, for which precise asymptotics for the max- $\ell_2$  margin was established in [31].

In Section 3.5.2, we prove a universality result of the following form: the value of the max- $\ell_1$ -margin remains the same (asymptotically) under two different settings where the distribution of the features entered in the boosting algorithm vary. To describe in detail, suppose the observed data  $\{x_i, y_i\}$  still arises from the data-generating distribution considered for our aforementioned point-by-point contributions. However, the features feeding to the Boosting Algorithm (and thus in calculating the margin) are more complicated than the raw features  $x_i$ 's. We consider two different kinds of boosting features—(i) features  $a_i$  that take the form of a random feature model  $a_i = \sigma(F^\top x_i)$  [79, 55, 75], (ii) features  $b_i = \mu_0 \mathbf{1} + \mu_1 F^\top x_i + \mu_2 z_i$ , where the constants  $\mu_0, \mu_1, \mu_2$  are calibrated appropriately to match moments of  $a_i$ 's and  $b_i$ 's. Here,  $F$  is a random matrix in  $\mathbb{R}^{p \times d}$  and  $z_i$  has i.i.d.  $\mathcal{N}(0, 1)$  entries, independent of everything else. In each case, the max- $\ell_1$ -margin is calculated using the formula  $\kappa_{n, \ell_1}(\{r_i, y_i\}_{1 \leq i \leq n}) := \max_{\|\theta\|_1 \leq 1} \min_{1 \leq i \leq n} y_i r_i^\top \theta$ , where  $r_i = a_i$  (resp.  $b_i$ ) in Case (i) (resp. Case (ii)). Section 3.5.2 establishes that, when  $p, d$  both scale linearly with  $n$ , the (scaled) max- $\ell_1$ -margin has the same limiting value under both settings.

The aforementioned result holds under certain assumptions on the random feature matrix  $F$  and the non-linearity  $\sigma(\cdot)$ . (see Section 3.5.2 for details). But note that, conditional on  $F$ ,  $b_i$  is Gaussian whereas  $a_i$  is not. This universality result suggests that the margin value is asymptotically insensitive, at least under some settings, to nuanced properties of the feature distribution. Thus, results that apply for the Gaussian case might be relevant for certain non-Gaussian feature distributions. We further validate this through empirical observations in Section 3.5.2. On the technical front, our universality result starts with a leave-one-out argument from [55]. However, [55] considered loss functions satisfying certain smoothness and strong-convexity assumptions, which are grossly violated in our setting. This leads to several technical challenges that we handle by establishing new analytic results (Section 3.5.2 and Appendix B.2).

**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 overparametrization ratio  $\psi \in \Psi \subset [0, 6]$ , and a data-generating process where the covariates  $x_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, I_p)$ , and the response  $y_i | x_i = +1$  with probability  $\sigma(x_i^\top \theta_\star)$  where  $\sigma(t) = 1/(1 + e^{-t})$ , and  $y_i | x_i = -1$  otherwise. Each coordinate of  $\theta_\star$  is drawn i.i.d. from a Gaussian  $\mathcal{N}(0, 1/p)$ . For each  $\psi \in \Psi$ , we generate multiple samples of size  $n = 400$ , and calculate the max- $\ell_1$ -margin by two

methods: (i) the numerical solution  $\kappa_{n,\ell_1}$  to the corresponding linear program (LP) in (1.3); the blue points in Figure 1(a) depict these values (appropriately scaled), and, (ii) the asymptotic value  $\kappa_\star(\psi, \mu)$  predicted by our analytic formula in Theorem 3.1; the red points labeled as CGMT in Figure 1(a) represent these values. Calculating our theoretical predictions involves solving a complex *non-linear system of equations* defined in (3.9). This involved computing integrals, which we approximate via Monte-Carlo sums (5000 samples). Figure 1(b) compares the corresponding out-of-sample prediction error: the blue points show the generalization error  $\mathbb{P}_{\mathbf{x}, \mathbf{y}}(\mathbf{y} \cdot \mathbf{x}^\top \hat{\theta}_{n,\ell_1} < 0)$ , when  $\hat{\theta}_{n,\ell_1}$  is calculated from the LP, whereas the red points depict the asymptotic value predicted by our theory (Theorem 3.2). In both 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 [24]. This is also evidenced in the plot—the max- $\ell_1$ -margin is positive (resp. zero) above (resp. below) this threshold, and as expected, our theory matches the numerics accurately above the threshold.

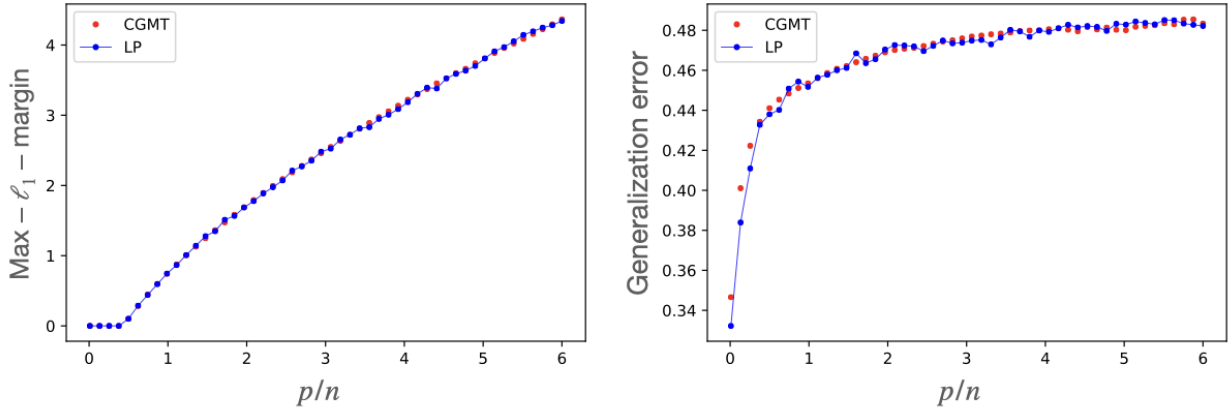


Figure 1:  $x$ -axis: Ratio  $p/n$ .  $y$ -axis: (a) Left: max- $\ell_1$ -margin, the blue points are obtained by solving the LP in (1.3) and averaging its solution over 10 independent simulation runs. The red points are obtained by numerically evaluating the formula in RHS of (3.2). (b) Right: Generalization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

**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. Unless otherwise mentioned, we consider a sequence of problems  $\{y(n), X(n), \theta_\star(n)\}_{n \geq 1}$ , such that  $y(n) \in \mathbb{R}^n$ ,  $\theta_\star(n) \in \mathbb{R}^{p(n)}$  and  $X(n) \in \mathbb{R}^{n \times p(n)}$ , where the  $i$ -th

row  $x_i \sim \mathcal{N}(0, \Lambda(n))$ , and the  $i$ -th entry of  $y(n)$  satisfies

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

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,  $p(n)/n \rightarrow \psi \in (0, \infty)$ . We require certain structural assumptions on the covariate distributions and the regression vector sequence that is described below. Conceptually, four factors determine the structure of the problem: overparametrization  $\psi$ , signal strength  $\rho$ , link function  $f$ , and a limiting measure  $\mu$  defined in Assumption 2. Later, Section 3.5 will investigate models beyond (2.1).

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

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

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

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

$$\rho(n) \rightarrow \rho \quad (2.3)$$

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

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

**Remark 2.1.** Note that Assumption 1 and (2.3) together imply that  $\sum_{j=1}^p \theta_\star(n)_j^2 = O(1)$ . If all the entries of  $\theta_\star$  are of the same order, this yields  $\theta_{\star,i} = O(1/\sqrt{p})$ . This also justifies why we include  $\sqrt{p}$  in the numerator of  $\bar{w}_i$ . The convergence in  $W_2$  equivalently means weak convergence and convergence of the second moments (see for instance, [103, 76]). In particular, this implies that  $\int w^2 \mu(d\lambda, dw) = 1$ .

**Assumption 3.** Finally, assume that

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

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

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

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 [30, 87, 64]. More recently, [24] derived the separability threshold  $\psi^*(\rho, f)$  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 [76]. To describe this phase transition threshold, consider the following bivariate function  $F_\kappa : \mathbb{R} \times \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}_{\geq 0}$  defined for any  $\kappa \geq 0$ ,

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

Then

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

As an example, recall that  $\psi^*(\rho, f) \approx 0.43$  in the setting of Figure 1. The above function  $F_\kappa : \mathbb{R} \times \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}_{\geq 0}$  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 [45, 44]. Suppose that each weak learner outputs a continuous decision  $X_{ij} = x_i[j] \in \mathbb{R}$  and  $y_i \in \{-1, +1\}$ . Let  $\Delta_n$  be the standard probability simplex given by  $\Delta_n := \{\mathbf{p} \in [0, 1]^n : \sum_{i=1}^n p_i = 1\}$ . Suppose  $Z = \mathbf{y} \circ X \in \mathbb{R}^{n \times p}$  denotes multiplying each element in the  $i$ -th row of  $X$  by  $y_i$ ,  $i \in [n]$ . At each step, AdaBoost adaptively chooses the best feature as follows:

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

### 3 Main Results

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

#### 3.1 Max- $\ell_1$ -margin and min- $\ell_1$ -norm interpolant

Recall the definition of the max- $\ell_1$ -margin from (1.3). We establish that  $\kappa_{n, \ell_1}$ , when appropriately scaled, converges almost surely to a limit that can be explicitly characterized in terms of  $\psi, \mu$  and  $f$ .

To describe this limit, consider the following function first introduced in [76]: for any  $(\psi, \kappa)$  pair that satisfies  $\psi > \psi^\downarrow(\kappa)$  (See Equation 3.12), define  $T : (\psi, \kappa) \rightarrow \mathbb{R}$  to be

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

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

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

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

where

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

The max- $\ell_1$ -margin was conjectured to be a central quantity for boosting [16]—Theorem 3.1 provides a precise high-dimensional characterization of this object under our data-generating scheme. For typical data instances, it is crucial to understand how such margin scales with the overparametrization, both theoretically and empirically, which is answered by the above Theorem. This limiting result will lead to precise characterizations of statistical and computational properties of *Boosting Algorithms* in 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  $\psi$ . Its dependence on the signal strength  $\rho$  and the distribution  $\mu$  is encoded through  $F_\kappa(\cdot, \cdot)$ , and the parameters  $c_1 \equiv c_1(\psi, \rho, \mu, \kappa)$ ,  $c_2 \equiv c_2(\psi, \rho, \mu, \kappa)$ ,  $s \equiv s(\psi, \rho, \mu, \kappa)$ , which appear in the definition of  $T(\psi, \kappa)$  (3.1).

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

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

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

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

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

Theorem 3.2 provides an exact quantification of the generalization behavior of the min- $\ell_1$ -norm interpolant under our data-generating scheme. Earlier works [83, 108] already characterized the long time and infinitesimal step size limit of AdaBoost on separable data. Later, Section 3.3 will establish a further precise connection between  $\hat{\theta}_{n,\ell_1}$  and the AdaBoost iterates (with suitably chosen learning rates). Informally, the AdaBoost iterates arrive arbitrarily close to the min- $\ell_1$ -norm interpolant, beyond a certain time threshold. Therefore, Theorem 3.2 provides two important contributions to the boosting literature, described as follows.

First, an open question was posed by Schapire et al. [89], Breiman [16] regarding which quantity truly governs the generalization performance of AdaBoost. Observe that in Theorem 3.2,  $\text{Err}_\star(\psi, \rho, \mu)$  crucially depends on  $\kappa_\star(\psi, \rho, \mu)$  (3.2) through the constants  $c_i^\star$ . Therefore, the asymptotic max- $\ell_1$ -margin precisely determines the generalization error. Since our result is asymptotically exact, Theorem 3.2 provides an answer to the question posed in [89, 16] under our assumed model. To contrast, the existing margin-based generalization upper bounds [89, 62] (that do not assume strong conditions on the data-generating distribution) scale as

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

In fact, note that the inverse of the  $y$ -axis in Figure 2 corresponds to the classical upper bound  $(\sqrt{n}\kappa_{n,\ell_1})^{-1}$  on the generalization error, as given by Eqn. (3.6), but this upper bound is vacuous in our setting (even overlooking the log factors) since it is worse than 0.5.

As a 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

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

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,  $\|\Pi_{(\Lambda^{1/2}\theta_\star)^\perp}(\Lambda^{1/2}\hat{\theta}_{n,\ell_1})\| \xrightarrow{\text{a.s.}} c_2^\star$ .

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

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

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 advances the literature on how the test error of boosting relates to the Bayes error [17, 59, 70, 107]: the optimality of Boosting (w.r.t. the optimal Bayes classifier) is entirely determined by the magnitude of  $(c_2^\star)^{-1}c_1^\star$ .

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

### 3.2 The non-linear system of equations

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

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

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

where

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

$$\mathcal{T} = \mathbf{prox}_s \left( \Lambda^{1/2} G + \psi^{-1/2} [\partial_1 F_\kappa(c_1, c_2) - c_1 c_2^{-1} \partial_2 F_\kappa(c_1, c_2)] \Lambda^{1/2} W \right),$$

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

Note that  $\Lambda$  denotes both the random variable in (3.9) and the covariance matrix in Assumption 1. Such overload of 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- $\ell_2$ -margin has received considerable attention in the past [76, 91, 48], however, (3.9) differs significantly from the equation system considered in case of the  $\ell_2$  geometry. This is natural, due to the intrinsic differences between the  $\ell_2$  and  $\ell_1$  geometries, and this also leads to significant additional technical challenges 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 Lasso under non-linear measurement models [100], and that of the MLE, LRT [96, 109] and convex regularized estimators [95, 86] for logistic regression.

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

$$\begin{aligned}\zeta &:= \left( \mathbb{E}_{(\Lambda, W) \sim \mu} |\Lambda^{-1/2} W| \right)^{-1} \\ \omega &:= \left( \mathbb{E}_{(\Lambda, W) \sim \mu} \left( W - \zeta \Lambda^{-1/2} \text{sign}(\zeta \Lambda^{-1/2} W) \right)^2 \right)^{1/2}\end{aligned}\quad (3.11)$$

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

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

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

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

Our proof for Proposition 3.1 adapts insights from [76] to the case of  $\ell_1$  geometry, however, the definition of  $\omega, \zeta$  in the threshold  $\psi^\downarrow(\kappa)$ , (3.12), differs from the case of  $\ell_2$  geometry. Now, it can be shown that  $F_\kappa(\cdot, \cdot)$  satisfies: (i)  $(\psi, \kappa) \mapsto T(\psi, \kappa)$  is continuous on its domain, (ii) for any fixed  $\kappa > 0$ ,  $T(\psi, \kappa)$  is strictly decreasing in  $\psi$ , (iii) for any fixed  $\psi > 0$ ,  $T(\psi, \kappa)$  is strictly increasing in  $\kappa$  ([76, Section B.5, Proposition 4.1]). Further, using the definition of  $\psi^\downarrow(\kappa)$ , and once again properties of  $F_\kappa(\cdot, \cdot)$ , one can establish that  $\lim_{\psi \rightarrow \infty} T(\psi, \kappa) < 0$ , whereas  $\lim_{\psi \downarrow \psi^\downarrow(\kappa)} T(\psi, \kappa) > 0$  and moreover,  $\lim_{\kappa \rightarrow \infty} T(\psi, \kappa) = \infty$ . Putting all of these together yields that the region  $\{(\psi, \kappa) : \psi > \psi^\downarrow(\kappa)\}$  contains the region  $\{(\psi, \kappa) : T(\psi, \kappa) = 0\}$ . This ensures (3.3) is well-defined, and that  $c_1^\star, c_2^\star, s^\star$  are unique. We defer to the Appendix for proof of Proposition 3.1.

### 3.3 Boosting in high dimensions

We turn our attention to the *Boosting Algorithm* described in Section 2. The path of boosting iterates was studied in infinite time and infinitesimal stepsize in [83, 108]. Here, we establish a sharp analysis of the number of iterations necessary for the AdaBoost iterates to approximately maximize the  $\ell_1$ -margin with arbitrary accuracy.

**Theorem 3.3.** *Under the assumptions of Theorem 3.1, with a suitably chosen learning rate (specified in 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*

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

the solution  $\hat{\theta}^t / \|\hat{\theta}^t\|_1$  forms  $(1 - \epsilon)$ -approximation to the Min- $\ell_1$ -Interpolated Classifier, that is, almost surely,

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

The above result is obtained by combining our Theorem 3.1 with a careful non-asymptotic analysis of AdaBoost allowing for an explicitly-specified learning rate, that builds upon existing works on margin maximization rates, rescaling and shrinkage techniques, and the mirror descent connections of AdaBoost (see [108, 98, 40, 52, 27, 57] and references cited therein). Together with Theorem 3.2, this result establishes a precise characterization of the computational and statistical behavior of AdaBoost for all iterations above the threshold  $T_\epsilon(n)$ , and notably complements the classical margin upper bounds [89, 62]. Thus, Theorem 3.3 reinforces a crucial conclusion from Section 3.1—the max- $\ell_1$ -margin is the key quantity governing the generalization error of AdaBoost in our setting.

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

**Remark 3.1.** A natural question may arise at this point: does the max- $\ell_1$ -margin studied here, when appropriately scaled, differ significantly from the  $\ell_2$ -margin [76]? Note that the rescaled  $\ell_1$ -margin is always larger than the  $\ell_2$ -margin, denoted by  $\kappa_{n, \ell_2}$ , since

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

A comparison of Figure 2 with [76, Fig. 1] shows that the range for the  $\ell_1$ -margin is roughly twice that for the  $\ell_2$  case, demonstrating that these behave differently, even after appropriate scaling.

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

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

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



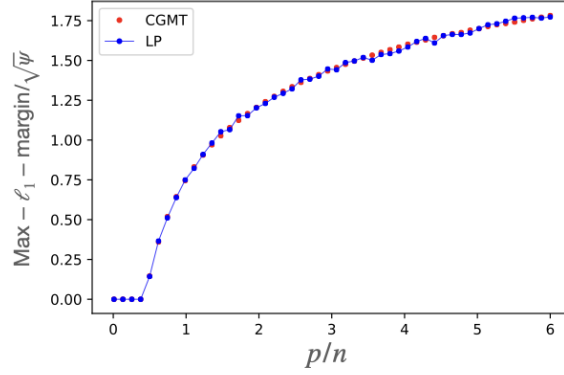


Figure 2:  $x$ -axis: varying ratio  $\psi := p/n$ .  $y$ -axis:  $\text{Max-}\ell_1\text{-margin}/\sqrt{\psi}$ . The setting is the same as in Figure 1. See Figure 1(a) for details on calculation of the blue and red points.

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

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

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

### 3.4 A new class of boosting algorithms

This section studies variants of AdaBoost that converge to the  $\text{max-}\ell_q\text{-margin}$  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  $\text{max-}\ell_q\text{-margin}$  and the  $\text{min-}\ell_q\text{-norm}$  interpolant beyond the case of  $q = 1$ . This complements the study of general  $\ell_q$  constraints, that was initiated by [83] (see also [52] and references therein). To this end, define the  $\text{max-}\ell_q\text{-margin}$  to be

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

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

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

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

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

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

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

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

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

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

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

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

With

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

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

define

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

Consider the system of equations

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

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

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

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

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

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

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

### 3.5 Robustness to assumptions

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

#### 3.5.1 Beyond independent covariates

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

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

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

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

where  $\Lambda \in \mathbb{R}^{p \times p}$  is a diagonal matrix. By the Bayes' formula, the conditional distribution of  $y_i|x_i$  can be captured through a logistic model, with  $\mathbb{P}(y_i = +1|x_i) = f\left(\log \frac{v}{1-v} + \langle \Lambda^{-1} \theta_\star, x_i \rangle\right)$  and  $f(t) = 1/(1 + e^{-t})$ . The covariate distribution obeys a mixture of Gaussians but the marginal covariance is given by  $\text{Cov}(x_i) = 4v(1-v)\theta_\star\theta_\star^\top + \Lambda$  (thus called the spiked covariance model). Compared to the diagonal covariance as in (2.1), the setting considered here therefore goes beyond independent covariates by introducing a rank-one spike to the diagonal covariance  $\Lambda$ .

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

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

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

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

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

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

Then, in the regime where the data is asymptotically linearly separable (see [31, Proposition 3.1] for the linear separability threshold for this problem), the max- $\ell_1$ -margin and min- $\ell_1$ -norm interpolant obey the limiting characterizations from Theorems 3.1-3.2, with the system of equations given by (3.27), and  $F_\kappa(c_1, c_2)$  substituted by (3.26). Note that [31] analyzed the max- $\ell_2$ -margin for a (misspecified) logistic model and the Gaussian mixture model (3.23)-(3.24) through a unified CGMT based analysis. Due to crucial differences between  $\ell_1$  and  $\ell_2$  geometries, the  $\ell_1$  case, (or for that matter, any  $\ell_q$  with  $q \neq 2$ ) does not follow directly from these results. We will elaborate on this point in Section 5.

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

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

with  $(y_i, m_i, \tilde{x}_i)$  independent of each other,  $m_i$  any random variable symmetric around zero,  $\tilde{x}_i \sim \mathcal{N}(0, \Lambda)$  with  $\Lambda$  diagonal. The observed data contains only  $(y_i, x_i)$  and thus, the  $m_i$ 's may be thought of as latent random variables. Note in this case,  $\text{Cov}(x_i) = 4v(1-v)\theta_\star\theta_\star^\top + \text{Var}(m_i)\tilde{\theta}\tilde{\theta}^\top + \Lambda$ , a rank-two perturbation of a diagonal covariance matrix. The aforementioned characterization can be naturally extended with appropriate analogues of (3.25)-(3.27). We assume that the Wasserstein-2

limit of the empirical distribution sequence  $\sum_{j=1}^p \delta_{(\lambda_i, \sqrt{p}\theta_\star^\top e_i, \sqrt{p}\tilde{\theta}^\top e_i)}/p$  exists, denote it by  $\tilde{\mu}$ , and let  $(\Lambda, h_\star, \tilde{h}, G) \sim \tilde{Q} = \tilde{\mu} \otimes \mathcal{N}(0, 1)$ . Define the following analogue of (3.26),

$$\tilde{F}_\kappa(c_1, c_2, c_3) = \sqrt{\mathbb{E}[(\kappa - c_1 - c_2\tilde{Z} - c_3M)_+^2]}, \quad (3.29)$$

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

$$\begin{aligned} c_1 &= \mathbb{E}_{(\Lambda, h_\star, \tilde{h}, G) \sim \tilde{Q}}[h_\star h_{\text{sol}}], & c_2^2 &= \mathbb{E}_{(\Lambda, h_\star, \tilde{h}, G) \sim \tilde{Q}}[(\Lambda^{1/2} h_{\text{sol}})^2], \\ c_3 &= \mathbb{E}_{(\Lambda, h_\star, \tilde{h}, G) \sim \tilde{Q}}[\tilde{h} h_{\text{sol}}], & 1 &= \mathbb{E}_{(\Lambda, h_\star, \tilde{h}, G) \sim \tilde{Q}}[|h_{\text{sol}}|], \end{aligned} \quad (3.30)$$

$$\text{where } h_{\text{sol}} = -\frac{\text{prox}_s(\Lambda^{1/2}G + \psi^{-1/2}(\partial_1 \tilde{F}_\kappa(c_1, c_2, c_3)h_\star + \partial_3 \tilde{F}_\kappa(c_1, c_2, c_3)\tilde{h}))}{\Lambda\psi^{-1/2}c_2^{-1}\partial_2 \tilde{F}_\kappa(c_1, c_2, c_3)}.$$

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

For both the aforementioned models, the boosting algorithm satisfies Theorem 3.3 with the respective limiting characterization of the max- $\ell_1$ -margin. A common theme across these settings is that the behavior of the margin and interpolant can be accurately characterized by a fixed point equation system, the solution to which possesses precise physical meanings (see (3.7) and the discussion thereafter). The form of the systems vary from one model to another; however, principles underlying its origin and key proof steps remain essentially the same (Section 5). Once again, this is the power of our theoretical analysis in the  $\ell_1$  case: we introduce a new uniform deviation argument with sufficient generality so that our proof can be adapted across several modeling schemes, as illustrated through this section.

To conclude this section, we showcase the numerical accuracy of our results for the rank-one spike case (3.23)-(3.24). The example is illustrated in Fig. 3. Here,  $\Lambda$  is always taken to be the identity matrix. The x-axis denotes the overparametrization ratio  $\psi = p(n)/n$ , y-axis the signal-to-noise ratio  $\rho = (\|\sqrt{p}\theta_\star\|^2/\text{Tr}(\Lambda))^{1/2}$ , and the color encodes the value of the max- $\ell_1$ -margin (top row) or prediction error of the corresponding min- $\ell_1$ -norm interpolant (bottom row) respectively. Thus, for each value on the y-axis, we choose a different signal  $\theta_\star$  so that the signal-to-noise ratio matches the given value of  $\rho$ . The left panel numerically solves the fixed-point equation (3.27) and presents the limits of the margin and prediction error from Theorems 3.1-3.2, obtained upon replacing (3.9) for the equation system in this rank-one spike case, (3.27). The right panel presents the max- $\ell_1$ -margin in finite samples, obtained by solving the LP (1.3), along with the corresponding prediction error, and these are averaged over two independent simulation runs. As Fig. 3 illustrates, the finite-sample results conform to our asymptotic characterization remarkably well. We defer further extensions to general feature covariance matrices not covered here or in Appendix B.1 for future work. Remark 5.2 (Section 5.1) explains the additional difficulty faced in such extensions.

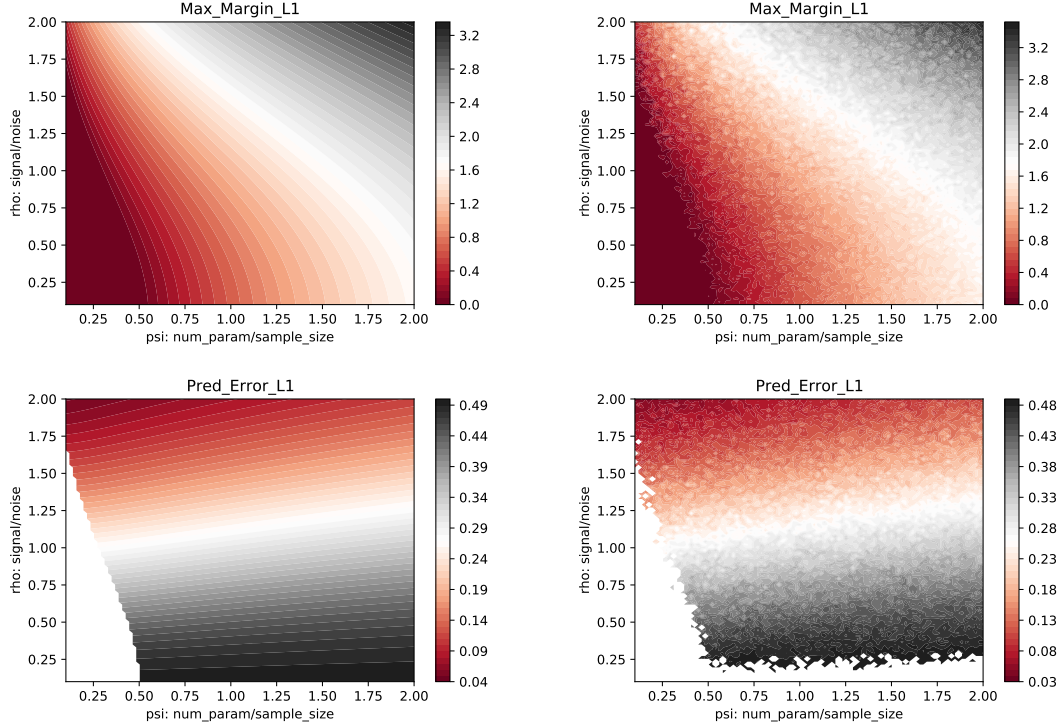


Figure 3:  $x$ -axis: Ratio  $p/n$ ,  $y$ -axis: Signal-to-noise ratio  $\rho = (\|\sqrt{p}\theta_\star\|^2/\text{Tr}(\Lambda))^{1/2}$ . The top row shows max- $\ell_1$ -margin and bottom row the prediction error of the corresponding interpolant. The left panel plots the limits of these objects, as characterized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

### 3.5.2 Beyond Gaussian covariates

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

This section will show that the max- $\ell_1$ -margin for the above non-linear random features model, in the asymptotic sense, equals that of an analogous Gaussian features model, conditioned on  $F$ . To be concrete, we show the asymptotic equivalence of max- $\ell_1$ -margin for two models: (i) random features  $a_i = \sigma(F^\top x_i) \in \mathbb{R}^d$ , and (ii) analogous Gaussian features  $b_i = \mu_0 \mathbf{1} + \mu_1 F^\top x_i + \mu_2 z_i \in \mathbb{R}^d$ , where



$z_i \sim \mathcal{N}(0, I_d)$ ,  $\mu_0 = \mathbb{E}[\sigma(Z)]$ ,  $\mu_1 = \mathbb{E}[Z\sigma(Z)]$ ,  $\mu_2 = \sqrt{\mathbb{E}(\sigma^2(Z)) - \mu_0^2 - \mu_1^2}$ , with  $Z \sim \mathcal{N}(0, 1)$  independent of everything else. Here  $\mu_0, \mu_1$  are top-two Hermite coefficients of  $\sigma(\cdot)$ , and  $\mu_2$  is the  $\ell_2$  norm of the remaining Hermite coefficients. The max- $\ell_1$ -margin under each model is calculated using  $\kappa_{n, \ell_1}(\{r_i, y_i\}_{1 \leq i \leq n}) := \max_{\|\theta\|_1 \leq 1} \min_{1 \leq i \leq n} y_i r_i^\top \theta$ , where  $r_i$  equals  $a_i$  or  $b_i$  depending on the model. We establish that the asymptotic value of the margin (scaled by  $\sqrt{p}$ ) remains the same irrespective of the choice of the features included in the calculation.

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

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

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

The above theorem asserts that, asymptotically, both the non-linear feature matrix  $A(n)$  and its Gaussian counterpart  $B(n)$  yield the same margin value. We next provide a brief outline of the proof. In Section 5.1, we mention that studying the limiting value of the margin is equivalent to studying whether  $\xi_{\psi, \kappa}^{(n, p)}(R) = \min_{\|\theta\|_1 \leq \sqrt{p}} \frac{1}{\sqrt{p}} \|(\kappa \mathbf{1} - (y \odot R)\theta)_+\|_2$  is strictly positive or not, where  $R$  denotes the feature matrix used in the margin definition. This is equivalent to studying  $\{\xi_{\psi, \kappa}^{(n, p)}(R)\}^2 = \min_{\|\tilde{\theta}\|_1 \leq p} \frac{1}{p} \sum_{i=1}^n (\kappa - \frac{1}{\sqrt{p}} y_i r_i^\top \tilde{\theta})_+^2$ , where we apply the change of variable  $\tilde{\theta} = \sqrt{p}\theta$ . Denote the Lagrange form for this problem with multiplier  $\lambda$  to be  $\Phi_n(R, \lambda)$ . We claim that to show (3.31), it suffices to show that for all  $\lambda$

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

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

$$\frac{1}{p} \min_{\|\tilde{\theta}\|_1 \leq p} \sup_{\lambda \geq 0} \sum_{i=1}^n (\kappa - \frac{1}{\sqrt{p}} y_i a_i^\top \tilde{\theta})_+^2 + \lambda \sum_{j=1}^p (|\tilde{\theta}_j| - 1) \quad (3.33)$$

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

To prove (3.32), we start with a leave-one-out argument adapted from [55], which in turn builds upon [37]. In [55], the authors prove that the training and generalization errors are asymptotically equivalent in a random features model and a corresponding linearized model, where the covariates have matching moments and are Gaussian conditional on the random features. However, [55]

defined the training error to be based on the objective function of a penalized empirical risk minimization problem, where the loss admits derivatives upto the third order and the regularizer is strongly convex. In our setting, neither of these properties hold, and this leads to several technical challenges. To handle these, we use a specific smoothing argument and develop several new analytic results (Appendix B.2).

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

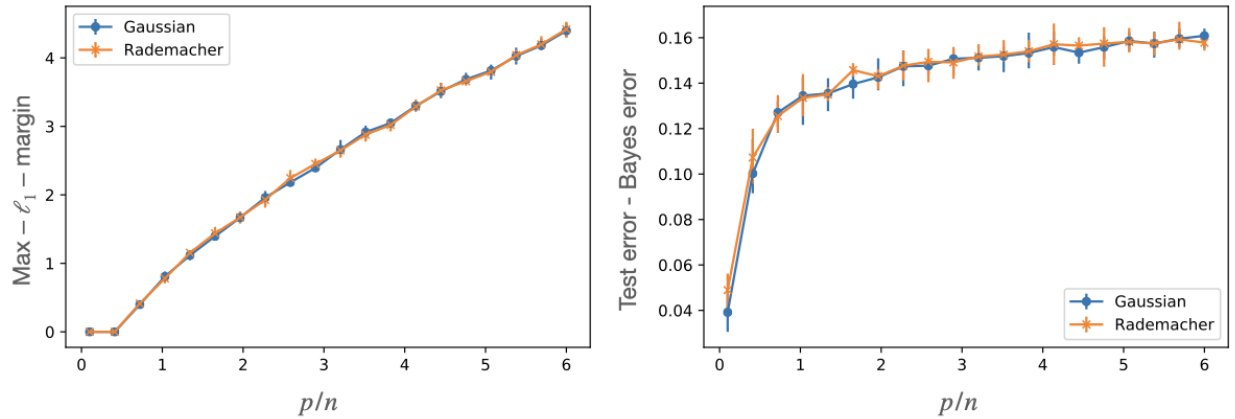


Figure 4:  $x$ -axis: Ratio  $p/n$ .  $y$ -axis: (Left subfigure) max- $\ell_1$ -margin, (Right subfigure) Test error minus the Bayes error. The figure has the same setting as in Figure 1, except the covariate distribution. Here, the observed design matrix has i.i.d. entries drawn either from a Rademacher distribution or a Gaussian with matching first and second moments. The figure demonstrates universality of the margin value and the test error across these settings.

### 3.5.3 Model Misspecification

Consider the following data generating process: denote  $\tilde{x}_i = (x_i^\top, z_i^\top)^\top$  where  $x_i \in \mathbb{R}^p$  and  $z_i \in \mathbb{R}^q$ , with  $x_i \sim \mathcal{N}(0, \Lambda_x)$  and  $z_i \sim \mathcal{N}(0, \Sigma_z)$  independent Gaussian vectors. Here we assume that  $\Lambda_x$  is a

diagonal matrix. Suppose that  $y$  arises from the following conditional distribution

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

The observed data contains  $n$  i.i.d. samples  $(x_i \in \mathbb{R}^p, y_i \in \mathbb{R}), 1 \leq i \leq n$ , 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

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

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

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

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

$$\begin{aligned} \tilde{F}_\kappa(c_1, c_2) &:= \left( \mathbb{E} \left[ (\kappa - c_1 Y Z_1 - c_2 Z_3)_+^2 \right] \right)^{\frac{1}{2}} \\ \text{where } &\begin{cases} Z_3 \perp (Y, Z_1, Z_2) \\ Z_i \stackrel{i.i.d.}{\sim} \mathcal{N}(0, 1), \quad i = 1, 2, 3 \\ \mathbb{P}(Y = +1|Z_1, Z_2) = 1 - \mathbb{P}(Y = -1|Z_1, Z_2) = f(\rho \cdot Z_1 + \gamma \cdot Z_2) \end{cases} \end{aligned} \quad (3.35)$$

Consider the regime where the observed data is asymptotically linearly separable, that is,  $\psi + \phi$  lies above the separability threshold for this problem. We do not describe the threshold here in detail, the interested reader may find its characterization in [31, Proposition 3.1]. Then the max- $\ell_1$ -margin and min- $\ell_1$ -norm interpolant, computed using the observed data  $\{(x_i, y_i)\}_{i=1}^n$  obey the same limiting characterizations as in Theorems 3.1-3.2, with the system of equations remaining the same as in (3.9), but with  $F_\kappa(c_1, c_2)$  substituted by the new function (3.35). Thus, the form of the equation system (3.9) once again remains unchanged, once we pin down the right analogue of  $F_\kappa(c_1, c_2)$  in this new setting.

## 4 Related Literature

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

**Boosting.** Since its introduction in [44, 45], there has been a vast and expansive literature on Boosting. [14] studied bias and variance of general arcing classifiers. A wonderful survey of early works on generalization performance of boosting, and comparisons to the optimal Bayes error can be found in [58]. Margin-based analyses were furthered in [81, 63, 80, 82]. For analysis of boosting algorithms based on smooth margin functions, see [85] and the references cited therein. Consistency properties were extensively studied in [71, 73, 72, 12]. 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 [46]. Subsequently, [47] developed a general gradient boosting framework. The rate of convergence of regularized

boosting classifiers was explored in [13], where the authors uncovered that some versions of boosting work especially well in high-dimensional logistic additive models.  $\ell_2$ -boosting, sparse boosting, twin boosting, and their properties in high dimensions were extensively studied in [21, 19, 22, 18, 20]. We remark that our setting is different in nature from this high-dimensional Boosting literature, where a notion of sparsity (often in  $\ell_1$  geometry) is typically assumed on the unknown parameter  $\theta_*$ . On the contrary, the  $\ell_1$  connection arises naturally in our setting, due to the nature of the AdaBoost/boosting algorithm. The rate of convergence of AdaBoost to the minimum of the exponential loss was investigated in [77]. Robust versions of boosting were proposed and extensively explored in [65]. In recent times, [41] 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 [49, 50], and is obtained by extending Gordon’s inequalities with the presence of convexity. The idea of merging these seemingly disparate threads dates back to [92, 93, 94], where it was used to analyze the performance of the constrained LASSO in high signal-to-noise ratio regimes. The seminal works [101, 99, 102] built and significantly extended on this idea to arrive at the CGMT, which was extremely useful for studying mean-squared errors of regularized M-estimators in high-dimensional linear models. As discussed earlier, [76] studied the asymptotic properties of the max- $\ell_2$ -margin in binary classification settings, building upon CGMT-based techniques, and furthered the work by [48]. In a similar setting, [31] 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 [54], to study high-dimensional regularized estimators in logistic regression [86], and to establish performance guarantees for PhaseMax [32]. 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- $\ell_1$ -norm interpolated classifier, which characterizes the limit of the Boosting solution on separable data. In recent years, min-norm interpolated solutions and their statistical properties have been extensively studied—see [9, 10, 66, 11, 53, 7, 68, 67, 23] for the regression problem, and [76, 31, 26] 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 [106, 9, 66]. To the best of our knowledge, the current paper is the first to provide sharp statistical results for interpolated classifiers induced by the  $\ell_1$  geometry (rather than the  $\ell_2$ ), which has been argued to be a more suitable geometry [5, 52, 35, 27, 4] 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 for Theorems 3.1 and 3.2

The proofs of Theorems 3.1 and 3.2 rely on the *Convex Gaussian Min-Max Theorem* (CGMT) [101, 99], which is a refinement of Gordon's classical Gaussian comparison inequality [50]. Our analysis is partially influenced by the seminal work of [76], which characterized the max- $\ell_2$ -margin using CGMT-based techniques. However, characterizing the asymptotics for the  $\ell_1$  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  $F_\kappa$ , which might be of standalone interest (we establish this in Lemma 5.1). Additionally, our analysis is general and extendable to the max- $\ell_q$ -margin case with  $1 \leq q \leq 2$ . 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

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

It is not hard to see that

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

Thus, to study the rescaled max- $\ell_1$ -margin, it suffices to examine the value of  $\xi_{\psi, \kappa}^{(n,p)}$ .

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

$$x_i^\top \theta_\star = z_i^\top \Lambda^{1/2} \theta_\star = \rho(n) \cdot z_i^\top w, \text{ where } w := \Lambda^{1/2} \theta_\star / \|\Lambda^{1/2} \theta_\star\|. \quad (5.3)$$

Using the fact that  $y \odot X = (y \odot Z) \Lambda^{1/2} \stackrel{d}{=} ((y \odot z) w^\top + Z \Pi_{w^\perp}) \Lambda^{1/2}$  (such a trick was first used in the literature in [100]), 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

$$\begin{aligned}\xi_{\psi, \kappa}^{(n,p)}(z, Z) &:= \min_{\|\theta\|_1 \leq \sqrt{p}} \max_{\|\lambda\|_2 \leq 1, \lambda \geq 0} \frac{1}{\sqrt{p}} \lambda^T \left( \kappa \mathbf{1} - (y \odot z) \langle w, \Lambda^{1/2} \theta \rangle \right) \\ &\quad - \frac{1}{\sqrt{p}} \lambda^T Z \Pi_{w^\perp} (\Lambda^{1/2} \theta).\end{aligned}\quad (5.4)$$

**Remark 5.1.** The rescaling by  $\sqrt{p}$  is required to ensure a well-defined limit for the max- $\ell_1$ -margin (in general, a rescaling by  $p^{1/q-1/2}$  is required for general  $\ell_q$  margin, as evidenced via Corollary 3.22, and this immediately shows that no rescaling is required for the  $\ell_2$  case [76]).

**Step 2: Reduction to Gordon's problem.** Due to the min-max form of (5.4), one can use Gordon's Gaussian comparison inequality [101, 99, 50] to further simplify the problem. To this end, introduce the following "de-coupled" optimization problem

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

where  $\mathcal{V} = (\kappa \mathbf{1} - (y \odot z) \langle w, \Lambda^{1/2} \theta \rangle - \tilde{z} \|\Pi_{w^\perp}(\Lambda^{1/2} \theta)\|_2)$ ,  $z, \tilde{z} \in \mathbb{R}^n$  and  $g \in \mathbb{R}^p$  are independent isotropic Gaussian vectors. By CGMT [101, Theorem 3] (see Theorem A.1 in the Appendix), we have

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

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

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

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

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  $\lambda = \text{diag}(\Lambda)$  denoting the vectorized  $\Lambda$ , we can express  $\hat{\xi}_{\psi,\kappa}^{(n,p)}(z, \tilde{z}, g)$  as the positive part of the following expression

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

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

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

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

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

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

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



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  $\hat{\xi}_{\psi,\kappa}^{(n,p)}(\lambda, w, g)$  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  $\ell_2$ -constraint  $\|\theta\|_2 \leq 1$  (the setting of [76]) and the  $\ell_1$ -constraint  $\|\theta\|_1 \leq \sqrt{p}$  (our setting),  $c_1$  is bounded in the sense  $|c_1| \leq M$  for all  $p(n), n$  and some constant  $M > 0$ ; for the  $\ell_1$  case, this follows by noting that

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

by Assumption 3. Turning to the second variable  $c_2$ , we see that under our  $\ell_1$ -constraint,  $c_2$  may potentially grow as  $\sqrt{p}$  whereas it remains bounded when the  $\ell_2$ -norm of  $\theta$  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  $c_2 \in [0, \infty)$  fail to deliver sharp results. To overcome this technical challenge, we discover a key self-normalization property of the partial derivatives of  $F_\kappa$  (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  $1 - n^{-2}$ ,*

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

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) (this finite  $n, p$  problem is not convex in  $\theta$ , therefore, the KKT conditions are merely necessary conditions in this case) converge uniformly, over an unbounded region for  $c_2$ , to the corresponding “fp equations obtained from the KKT conditions for (5.10)”. (The KKT conditions are both necessary and sufficient in this case. See Appendix A.2 for details.) The convergence here 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  $(c_1^*, c_2^*, s^*)$  of the fp equations from (5.10). See Appendix A.4 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  $s^*$  and  $F_\kappa, \partial_i F_\kappa, i = 1, 2$  evaluated at  $(c_1^*, c_2^*)$ . 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  $\ell_1$  geometry. A critical, and perhaps surprising, consequence of our uniform deviation results is a localization property: any optimizer of (5.9) possesses finite  $\ell_2$ -norm.

#### Step 4: Fixed point equations and final step.

By standard analysis arguments (see Appendix A.4), the KKT conditions for the optimization problem (5.10) can be expressed as

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

From properties of the proximal mapping operator, the KKT conditions suggest that the solution must satisfy (see Appendix A.4 for a derivation of this claim, and the proof of uniqueness of the solution)

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

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 (3.9), 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) explains how  $s^*$  from Section 3.1 (the third component in the solution to our system of equations (3.9)) corresponds to Lagrange multipliers induced by the  $\ell_1$  constraint in (5.13).

**Remark 5.2.** As described in Section 3.5.1, the above proof path can accommodate a broad class of feature covariance matrices that are finite-rank perturbations of a diagonal (see Appendix B.1 for the general case extension). However, further extensions beyond this class poses an additional challenge: a crucial step in our proofs lies in establishing a large sample limit of a finite dimensional optimization problem (e.g. that in (5.9) or (B.3)). Here, the limit is described in terms of an optimization problem over  $\{h: \mathbb{R}^3 \rightarrow \mathbb{R}, h \in \mathcal{L}^2(\mathcal{Q})\}$ , where  $\mathcal{Q} = \mu \otimes \mathcal{N}(0, 1)$  ((5.10) or (B.4)). In the case of  $\ell_1$  geometry (and other  $\ell_q$  geometry for  $q \neq 2$ ), going over to this function space is feasible for a diagonal covariance matrix or non-diagonal matrices with a special structure. Instead, if  $\Lambda$  were a general non-diagonal matrix (not included in the classes considered in Section 3.5.1 and Appendix B.1), this leads to an added challenge. The finite sample optimization problem still retains a similar form as in (5.9), however, it is unclear how to express its limit in a convenient way and handle the terms  $\Lambda^{1/2}\theta$ . Given that the  $\ell_1$  theory requires several technical contributions over prior works, as described in this section, and Sections 3.5.1 and 3.5.2, we defer the case of more general covariance matrices for future work. We comment that the challenge faced here is similar in spirit to that seen in the context of the Lasso under arbitrary covariance, when studied under our proportional asymptotics regime. Here, one can be establish that this problem is asymptotically equivalent to a Gaussian sequence model with correlated errors. Now, this latter model is complicated and extracting neat characterizations from this equivalent problem remains quite a challenge (see for instance [25, 2, 56] for some progress in this direction).

## 5.2 Proofs of Theorems 3.3 and Corollary 3.1

[108] employs a re-scaling technique to establish that Boosting with infinitesimal stepsize agrees with the  $\min\text{-}\ell_1$ -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 use a simple yet general analysis of Boosting as a special instance of Mirror Descent (the connection between AdaBoost and mirror descent is well-known and it is infeasible to provide a complete list of references establishing and utilizing this. We refer the interested reader to Sections 1 and 3.3 for a partial list of related works) in conjunction with the re-scaling technique [108] and the shrinkage technique [98] (note this latter work also develops a  $1/\sqrt{t}$  margin maximization rate). Our analysis is similar in spirit to [29], but with different executions. One benefit of our analysis is that it is easily generalizable to a variant of boosting algorithm that maximizes  $\ell_q$  margin with  $q \geq 1$ .

**Proposition 5.1.** *Consider the Boosting Algorithm stated in Section 2. 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*

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

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

**Corollary 5.1** (Boosting converges to max- $\ell_1$ -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  $\theta_T$  converge to the max- $\ell_1$ -margin Direction in the following sense: for any  $0 < \epsilon < 1$ ,*

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

*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 3.3 and Corollary 3.1, we choose  $M(\delta) = \sqrt{(3 + \delta) \log(np)}$  for arbitrarily small  $\delta > 0$  and recall that  $\sqrt{p} \kappa_{n,\ell_1} \xrightarrow{\text{a.s.}} \kappa_\star(\psi, \rho, \mu)$ . 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  $\lim_{\delta \rightarrow 0} 2M^2(\delta)/\log n = 12$ .

## 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- $\ell_1$ -margin, the min- $\ell_1$ -norm interpolant, and a sharp analysis of the time necessary for AdaBoost to approximate this interpolant arbitrarily well. In doing so, this work identifies the exact quantities that govern the generalization behavior of AdaBoost for a class of data-generation models, and the relationship between this test error and the optimal Bayes error. On the optimization front, we further uncover how overparametrization leads to faster optimization. The proposed theory demonstrates commendable finite sample behavior, applies for

a broad class of statistical models, and is empirically robust to violations of certain assumptions. Natural variants of AdaBoost that correspond to  $\max\text{-}\ell_q$ -margins for  $q > 1$ , are further analyzed.

We conclude with a couple of directions of future research: it would be of interest (a) to rigorously characterize analogous properties of AdaBoost for covariate distributions with arbitrary correlations; this is a particularly challenging task for general  $\ell_q$  geometry when  $q \neq 2$ , as explained in Remark (5.2), and (c) to complement such characterizations via data-driven schemes for estimating the parameters  $c_1^*, c_2^*$  that govern properties of the  $\ell_1$  margin and interpolant, as well as the generalization performance of AdaBoost. Such estimation schemes are expected to be useful for providing recommendations regarding algorithm choice to practitioners.

## 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] Ayed M Alrashdi, Houssein Sifaou, Abba Kammoun, Mohamed-Slim Alouini, and Tareq Y Al-Naffouri. Precise error analysis of the lasso under correlated designs. *arXiv preprint arXiv:2008.13033*, 2020.
- [3] 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.
- [4] Ehsan Amid and Manfred K Warmuth. Winnowing with gradient descent. *Proceedings of Machine Learning Research* vol, 125:1–20, 2020.
- [5] Francis Bach. Breaking the curse of dimensionality with convex neural networks. *The Journal of Machine Learning Research*, 18(1):629–681, 2017.
- [6] Peter L Bartlett and Mikhail Traskin. Adaboost is consistent. *Journal of Machine Learning Research*, 8(Oct):2347–2368, 2007.
- [7] Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *arXiv preprint arXiv:1906.11300*, 2019.
- [8] 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.
- [9] Mikhail Belkin, Siyuan Ma, and Soumik Mandal. To understand deep learning we need to understand kernel learning. *arXiv preprint arXiv:1802.01396*, 2018.
- [10] Mikhail Belkin, Alexander Rakhlin, and Alexandre B Tsybakov. Does data interpolation contradict statistical optimality? *arXiv preprint arXiv:1806.09471*, 2018.
- [11] Mikhail Belkin, Daniel Hsu, and Ji Xu. Two models of double descent for weak features. *arXiv preprint arXiv:1903.07571*, 2019.
- [12] 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.

- [13] 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.
- [14] Leo Breiman. Arcing classifiers. *Annals of Statistics*, 26:123–40, 1996.
- [15] Leo Breiman. Bias, variance, and arcing classifiers. Technical report, Tech. Rep. 460, Statistics Department, University of California, Berkeley . . . , 1996.
- [16] Leo Breiman. Prediction games and arcing algorithms. *Neural computation*, 11(7):1493–1517, 1999.
- [17] Leo Breiman. Population theory for boosting ensembles. *The Annals of Statistics*, 32(1):1–11, 2004.
- [18] Peter Bühlmann. Boosting for high-dimensional linear models. *The Annals of Statistics*, 34(2):559–583, 2006.
- [19] Peter Bühlmann and Torsten Hothorn. Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science*, 22(4):477–505, 2007.
- [20] Peter Bühlmann and Torsten Hothorn. Twin boosting: improved feature selection and prediction. *Statistics and Computing*, 20(2):119–138, 2010.
- [21] 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.
- [22] Peter Bühlmann and Bin Yu. Sparse boosting. *Journal of Machine Learning Research*, 7(Jun):1001–1024, 2006.
- [23] Florentina Bunea, Seth Strimas-Mackey, and Marten Wegkamp. Interpolation under latent factor regression models. *arXiv preprint arXiv:2002.02525*, 2020.
- [24] Emmanuel J Candès and Pragma 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.
- [25] Michael Celentano, Andrea Montanari, and Yuting Wei. The lasso with general gaussian designs with applications to hypothesis testing. *arXiv preprint arXiv:2007.13716*, 2020.
- [26] Niladri S Chatterji and Philip M Long. Finite-sample analysis of interpolating linear classifiers in the overparameterized regime. *arXiv preprint arXiv:2004.12019*, 2020.
- [27] Lenaïc 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.
- [28] Alexandra Chouldechova and Aaron Roth. The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810*, 2018.
- [29] Michael Collins, Robert E Schapire, and Yoram Singer. Logistic regression, adaboost and bregman distances. *Machine Learning*, 48(1-3):253–285, 2002.

- [30] 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.
- [31] 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.
- [32] 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.
- [33] 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.
- [34] David L Donoho, Arian Maleki, and Andrea Montanari. Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences*, 106(45):18914–18919, 2009.
- [35] Xialiang Dou and Tengyuan Liang. Training neural networks as learning data-adaptive kernels: Provable representation and approximation benefits. *Journal of the American Statistical Association*, 0(0):1–14, 2020. doi: 10.1080/01621459.2020.1745812.
- [36] Harris Drucker and Corinna Cortes. Boosting decision trees. In *Advances in neural information processing systems*, pages 479–485, 1996.
- [37] Nouredine 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.
- [38] Nouredine 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.
- [39] Oliver Y Feng, Ramji Venkataramanan, Cynthia Rush, and Richard J Samworth. A unifying tutorial on approximate message passing. *arXiv preprint arXiv:2105.02180*, 2021.
- [40] Robert M Freund, Paul Grigas, and Rahul Mazumder. Adaboost and forward stagewise regression are first-order convex optimization methods. *arXiv preprint arXiv:1307.1192*, 2013.
- [41] 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.
- [42] Yoav Freund. Boosting a weak learning algorithm by majority. *Information and computation*, 121(2):256–285, 1995.



- [43] 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.
- [44] 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.
- [45] Yoav Freund and Robert E Schapire. Experiments with a new boosting algorithm. In *icml*, volume 96, pages 148–156. Citeseer, 1996.
- [46] 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.
- [47] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.
- [48] Elizabeth Gardner. The space of interactions in neural network models. *Journal of physics A: Mathematical and general*, 21(1):257, 1988.
- [49] Yehoram Gordon. Some inequalities for gaussian processes and applications. *Israel Journal of Mathematics*, 50(4):265–289, 1985.
- [50] 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.
- [51] Adam J Grove and Dale Schuurmans. Boosting in the limit: Maximizing the margin of learned ensembles. In *AAAI/IAAI*, pages 692–699, 1998.
- [52] Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. *arXiv preprint arXiv:1802.08246*, 2018.
- [53] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*, 2019.
- [54] 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.
- [55] Hong Hu and Yue M Lu. Universality laws for high-dimensional learning with random features. *arXiv preprint arXiv:2009.07669*, 2020.
- [56] Hanwen Huang. Lasso risk and phase transition under dependence. *arXiv preprint arXiv:2103.16035*, 2021.
- [57] Ziwei Ji and Matus Telgarsky. Characterizing the implicit bias via a primal-dual analysis. In *Algorithmic Learning Theory*, pages 772–804. PMLR, 2021.

- [58] 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.
- [59] Wenxin Jiang. Process consistency for adaboost. *The Annals of Statistics*, 32(1):13–29, 2004.
- [60] 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.
- [61] Vladimir Koltchinskii and Olexandra Beznosova. Exponential convergence rates in classification. In *International Conference on Computational Learning Theory*, pages 295–307. Springer, 2005.
- [62] 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.
- [63] 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.
- [64] Emmanuel Lesaffre and Adelin Albert. Partial separation in logistic discrimination. *Journal of the Royal Statistical Society: Series B (Methodological)*, 51(1):109–116, 1989.
- [65] 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.
- [66] 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.
- [67] Tengyuan Liang and Hai Tran-Bach. Mehler’s formula, branching process, and compositional kernels of deep neural networks. *Journal of the American Statistical Association*, 0(0):1–14, 2021. doi: 10.1080/01621459.2020.1853547.
- [68] Tengyuan Liang, Alexander Rakhlin, and Xiyu Zhai. On the multiple descent of minimum-norm interpolants and restricted lower isometry of kernels. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of 33rd Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 2683–2711. PMLR, July 2020.
- [69] Zachary C Lipton. The mythos of model interpretability. *Queue*, 16(3):31–57, 2018.
- [70] 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.
- [71] Shie Mannor and Ron Meir. Geometric bounds for generalization in boosting. In *International Conference on Computational Learning Theory*, pages 461–472. Springer, 2001.
- [72] Shie Mannor and Ron Meir. On the existence of linear weak learners and applications to boosting. *Machine Learning*, 48(1-3):219–251, 2002.

- [73] 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.
- [74] Llew Mason, Jonathan Baxter, Peter L Bartlett, and Marcus R Freen. Boosting algorithms as gradient descent. In *Advances in neural information processing systems*, pages 512–518, 2000.
- [75] Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 2019.
- [76] 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.
- [77] 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.
- [78] 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.
- [79] Ali Rahimi and Benjamin Recht. Weighted Sums of Random Kitchen Sinks: Replacing minimization with randomization in learning. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 1313–1320. Curran Associates, Inc., 2009.
- [80] Gunnar Rätsch and Manfred K Warmuth. Efficient margin maximizing with boosting. *Journal of Machine Learning Research*, 6(Dec):2131–2152, 2005.
- [81] Gunnar Rätsch, Takashi Onoda, and K-R Müller. Soft margins for adaboost. *Machine learning*, 42(3):287–320, 2001.
- [82] 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.
- [83] 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.
- [84] 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.
- [85] 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.
- [86] 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.

- [87] 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.
- [88] Robert E Schapire. The strength of weak learnability. *Machine learning*, 5(2):197–227, 1990.
- [89] 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.
- [90] Shai Shalev-Shwartz and Yoram Singer. On the equivalence of weak learnability and linear separability: New relaxations and efficient boosting algorithms. *Machine learning*, 80(2): 141–163, 2010.
- [91] Mariya Shcherbina and Brunello Tirozzi. Rigorous solution of the gardner problem. *Communications in mathematical physics*, 234(3):383–422, 2003.
- [92] Mihailo Stojnic. A framework to characterize performance of lasso algorithms. *arXiv preprint arXiv:1303.7291*, 2013.
- [93] Mihailo Stojnic. Meshes that trap random subspaces. *arXiv preprint arXiv:1304.0003*, 2013.
- [94] Mihailo Stojnic. Upper-bounding l1-optimization weak thresholds. *available at arXiv*, 2013.
- [95] Pragya Sur. A modern maximum-likelihood theory for high-dimensional logistic regression. [url.stanford.edu/jw604jq1260](http://url.stanford.edu/jw604jq1260), Ph.D. thesis, Stanford University, 2019.
- [96] 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.
- [97] 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.
- [98] Matus Telgarsky. Margins, shrinkage, and boosting. *arXiv preprint arXiv:1303.4172*, 2013.
- [99] Christos Thrampoulidis, Samet Oymak, and Babak Hassibi. The gaussian min-max theorem in the presence of convexity. *arXiv preprint arXiv:1408.4837*, 2014.
- [100] Christos Thrampoulidis, Ehsan Abbasi, and Babak Hassibi. Lasso with non-linear measurements is equivalent to one with linear measurements. *Advances in Neural Information Processing Systems*, 28:3420–3428, 2015.
- [101] 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.
- [102] 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.

- [103] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.
- [104] Shuaiwen Wang, Haolei Weng, and Arian Maleki. Which bridge estimator is the best for variable selection? *The Annals of Statistics*, 48(5):2791–2823, 2020.
- [105] Adrian Weller. Transparency: motivations and challenges. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pages 23–40. Springer, 2019.
- [106] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.
- [107] Tong Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics*, pages 56–85, 2004.
- [108] Tong Zhang and Bin Yu. Boosting with early stopping: Convergence and consistency. *The Annals of Statistics*, 33(4):1538–1579, 2005.
- [109] 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 Main Proofs

### A.1 The Convex Gaussian Min-Max Theorem

For the convenience of the readers, we state the convex Gaussian min-max theorem below [101, Theorem 4] (see also [50])

**Theorem A.1.** *Let  $\Omega_1 \subset \mathbb{R}^n, \Omega_2 \subset \mathbb{R}^p$  be two compact sets and let  $U : \Omega_1 \times \Omega_2 \rightarrow \mathbb{R}$  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  $t \in \mathbb{R}$ ,

$$\mathbb{P}(V_1(Z) \leq t) \leq 2\mathbb{P}(V_2(g, h) \leq t) .$$

2. Suppose  $\Omega_1$  and  $\Omega_2$  are both convex, and  $U$  is convex-concave in  $(w_1, w_2)$ . Then, for all  $t \in \mathbb{R}$ ,

$$\mathbb{P}(V_1(Z) \geq t) \leq 2\mathbb{P}(V_2(g, h) \geq t) .$$

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

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

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

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

$$V_1^{(\infty, \infty)}(c_1, c_2, s) := c_1 +$$

$$\mathbf{E}_{(\Lambda, W, G) \sim \mathcal{Q}_\infty} \left( \frac{\Lambda^{-1/2} W \cdot \mathbf{prox}_s \left( \Lambda^{1/2} \Pi_{W^\perp}(G) + \psi^{-1/2} [\partial_1 F_\kappa(c_1, c_2) - c_1 c_2^{-1} \partial_2 F_\kappa(c_1, c_2)] \Lambda^{1/2} W \right)}{\psi^{-1/2} c_2^{-1} \partial_2 F_\kappa(c_1, c_2)} \right)$$

$$V_2^{(\infty, \infty)}(c_1, c_2, s) := c_1^2 + c_2^2 -$$

$$\mathbf{E}_{(\Lambda, W, G) \sim \mathcal{Q}_\infty} \left( \frac{\Lambda^{-1/2} \mathbf{prox}_s \left( \Lambda^{1/2} \Pi_{W^\perp}(G) + \psi^{-1/2} [\partial_1 F_\kappa(c_1, c_2) - c_1 c_2^{-1} \partial_2 F_\kappa(c_1, c_2)] \Lambda^{1/2} W \right)}{\psi^{-1/2} c_2^{-1} \partial_2 F_\kappa(c_1, c_2)} \right)^2 \quad (\text{A.2})$$

$$V_3^{(\infty, \infty)}(c_1, c_2, s) :=$$

$$1 - \mathbf{E}_{(\Lambda, W, G) \sim \mathcal{Q}_\infty} \left| \frac{\Lambda^{-1} \mathbf{prox}_s \left( \Lambda^{1/2} G + \psi^{-1/2} [\partial_1 F_\kappa(c_1, c_2) - c_1 c_2^{-1} \partial_2 F_\kappa(c_1, c_2)] \Lambda^{1/2} W \right)}{\psi^{-1/2} c_2^{-1} \partial_2 F_\kappa(c_1, c_2)} \right| ,$$



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  $(\psi, \kappa)$  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 (3.9). 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*

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

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

$$(c_1, c_2, s) \in \mathcal{B}\left((c_1^\star, c_2^\star, s^\star), \epsilon\right). \quad (\text{A.4})$$

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

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

and define

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

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

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

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

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

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

and the corresponding infinite-dimensional optimization problem given by

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

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

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

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

*Proof of Proposition A.1.* To begin with, recall the KKT conditions (A.78) and its consequences (A.87)-(A.88), 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  $\theta$  (due to  $\widehat{F}$ ). Nonetheless, for any  $\theta$  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  $\theta/\sqrt{p}, \mathcal{Q}_p, \widehat{F}_\kappa$  replacing  $h, \mathcal{Q}_\infty, F_\kappa$ , we obtain the finite sample versions

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

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

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

In the following, we will prove the above claims.

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

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

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

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

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

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

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

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

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

$$\text{Eqn. (A.15)} \leq (c_2 \vee 1)^{-1} \left\{ \frac{\|\Lambda^{-1/2} G\|_{L_2(\mathcal{Q}_p)} \|\Lambda^{1/2} W\|_{L_2(\mathcal{Q}_p)} + \|W\|_{L_2(\mathcal{Q}_p)}^2 |\hat{C}^\uparrow|}{|\hat{C}^\downarrow C^\downarrow|} |\hat{C}^\downarrow - C^\downarrow| + \frac{\|W\|_{L_2(\mathcal{Q}_p)}^2}{|C^\downarrow|} |\hat{C}^\uparrow - C^\uparrow| \right\}. \quad (\text{A.18})$$

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

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

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  $|c_1| \leq M, 0 < c_2 \leq M, s > 0$

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

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

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

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

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

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

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

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

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

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

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

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)

$$(c_2 \vee 1)^{-1} |V_1^{(\infty,p)}(c_1, c_2, s) - V_1^{(\infty,\infty)}(c_1, c_2, s)| \quad (\text{A.24})$$

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

where

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

Since  $\mathcal{Q}_p \xrightarrow{W_2} \mathcal{Q}_\infty$ , by Theorem 2.7 and Proposition 2.4 in [3], we know that (1) for any function  $g$  that grows at most quadratically,

$$\sup_{\Lambda, W, G} \frac{|g(\Lambda, W, G)|}{1 + \|(\Lambda, W, G)\|_2^2} < \infty, \quad (\text{A.27})$$

$$\lim_{p \rightarrow \infty} \left| \left( \mathbf{E}_{(\Lambda, W, G) \sim \mathcal{Q}_p} - \mathbf{E}_{(\Lambda, W, G) \sim \mathcal{Q}_\infty} \right) g(\Lambda, W, G) \right| = 0, \quad (\text{A.28})$$

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

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

Here  $\mathbb{R}^3 \setminus \mathcal{B}_{R_\epsilon}$  denotes the complement of a ball of radius  $R_\epsilon$  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  $|c_1| \leq M, c_2 > 0, s > 0$ . Observe that

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

Further, for all  $|c_1| \leq M, 0 \leq c_2 \leq M, s \geq 0$ , uniformly for  $\Lambda, W, G$  (recall that  $\Lambda, W$  has bounded domain)

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

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

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

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

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

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  $R_\epsilon$  such that (A.29) holds true. Therefore

$$\begin{aligned} & \left| \int g_{c_1, c_2, s} d\mathcal{Q}_p - \int g_{c_1, c_2, s} d\mathcal{Q}_\infty \right| \\ & \leq \left| \int_{B_{R_\epsilon}} g_{c_1, c_2, s} (d\mathcal{Q}_p - d\mathcal{Q}_\infty) \right| + \left| \int_{\mathbb{R}^3 \setminus B_{R_\epsilon}} g_{c_1, c_2, s} (d\mathcal{Q}_p - d\mathcal{Q}_\infty) \right| \\ & \leq \left| \int_{B_{R_\epsilon}} g_{c_1, c_2, s} (d\mathcal{Q}_p - d\mathcal{Q}_\infty) \right| + \left| \int_{\mathbb{R}^3 \setminus B_{R_\epsilon}} g_{c_1, c_2, s} d\mathcal{Q}_p \right| + \left| \int_{\mathbb{R}^3 \setminus B_{R_\epsilon}} g_{c_1, c_2, s} d\mathcal{Q}_\infty \right| , \\ & \leq \left| \int_{B_{R_\epsilon}} g_{c_1, c_2, s} (d\mathcal{Q}_p - d\mathcal{Q}_\infty) \right| + 2K\epsilon \end{aligned} \quad (\text{A.33})$$

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

$$\left| \int_{\mathbb{R}^3 \setminus B_{R_\epsilon}} g_{c_1, c_2, s} d\mathcal{Q}_p \right| \leq K \int_{\mathbb{R}^3 \setminus B_{R_\epsilon}} \|(\Lambda, W, G)\|^2 d\mathcal{Q}_p \leq K\epsilon . \quad (\text{A.34})$$

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

$$\left| \int_{B_{R_\epsilon}} g_{c_1, c_2, s} (d\mathcal{Q}_p - d\mathcal{Q}_\infty) \right| \leq L_{R_\epsilon} W_1(\mathcal{Q}_p, \mathcal{Q}_\infty) \leq L_{R_\epsilon} W_2(\mathcal{Q}_p, \mathcal{Q}_\infty) . \quad (\text{A.35})$$

Now we have proved that for

$$\sup_{|c_1| \leq M, c_2 > 0, s > 0} \left| \int g_{c_1, c_2, s} d\mathcal{Q}_p - \int g_{c_1, c_2, s} d\mathcal{Q}_\infty \right| \leq L_{R_\epsilon} W_2(\mathcal{Q}_p, \mathcal{Q}_\infty) + 2K\epsilon, \quad (\text{A.36})$$

$$\lim_{p \rightarrow \infty} \sup_{|c_1| \leq M, c_2 > 0, s > 0} \left| \int g_{c_1, c_2, s} d\mathcal{Q}_p - \int g_{c_1, c_2, s} d\mathcal{Q}_\infty \right| \leq 2K\epsilon. \quad (\text{A.37})$$

By the fact that  $\epsilon$  can take an arbitrarily small value, we have proved

$$\lim_{n \rightarrow \infty, p(n)/n = \psi} \sup_{|c_1| \leq M, c_2 > 0, s > 0} (c_2 \vee 1)^{-1} |V_1^{(\infty, p)}(c_1, c_2, s) - V_1^{(\infty, \infty)}(c_1, c_2, s)| = 0, \quad (\text{A.38})$$

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

$$\lim_{n \rightarrow \infty, p(n)/n = \psi} \sup_{|c_1| \leq M, c_2 > 0, s > 0} (c_2 \vee 1)^{-1} |V_1^{(n, p)}(c_1, c_2, s) - V_1^{(\infty, p)}(c_1, c_2, s)| = 0. \quad (\text{A.39})$$

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

$$\lim_{n \rightarrow \infty, p(n)/n = \psi} \sup_{|c_1| \leq M, c_2 > 0, s > 0} (c_2 \vee 1)^{-1} |V_1^{(n, p)}(c_1, c_2, s) - V_1^{(\infty, p)}(c_1, c_2, s)| = 0, \quad a.s. \quad (\text{A.40})$$

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

$$\begin{aligned} & |V_2^{(n, p)}(c_1, c_2, s) - V_2^{(\infty, \infty)}(c_1, c_2, s)| \\ & \leq |V_2^{(n, p)}(c_1, c_2, s) - V_2^{(\infty, p)}(c_1, c_2, s)| + |V_2^{(\infty, p)}(c_1, c_2, s) - V_2^{(\infty, \infty)}(c_1, c_2, s)|. \end{aligned} \quad (\text{A.41})$$

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

$$(c_2 \vee 1)^{-2} |V_2^{(n, p)}(c_1, c_2, s) - V_2^{(\infty, p)}(c_1, c_2, s)| \quad (\text{A.42})$$

$$\begin{aligned} & \leq (c_2 \vee 1)^{-2} \mathbf{E}_{(\Lambda, W, G) \sim \mathcal{Q}_p} \left\{ \left( \frac{|\Pi_{W^\perp}(G)| + |\hat{C}^\uparrow| \|W\|}{|\hat{C}^\downarrow|} + \frac{|\Pi_{W^\perp}(G)| + |C^\uparrow| \|W\|}{|C^\downarrow|} \right) \right. \\ & \quad \left. \times \left( \frac{|\Pi_{W^\perp}(G)| + |\hat{C}^\uparrow| \|W\|}{|\hat{C}^\downarrow| |C^\downarrow|} |\hat{C}^\downarrow - C^\downarrow| + \frac{|W|}{|C^\downarrow|} |\hat{C}^\uparrow - C^\uparrow| \right) \right\}. \end{aligned} \quad (\text{A.43})$$

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

$$|c_2 \hat{C}^\downarrow - c_2 C^\downarrow|, |c_2 \hat{C}^\uparrow - c_2 C^\uparrow| \lesssim \frac{\log n}{\sqrt{n}}. \quad (\text{A.44})$$

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

$$|c_2 \hat{C}^\downarrow - c_2 C^\downarrow|, |\hat{C}^\uparrow - C^\uparrow| \lesssim \frac{\log n}{\sqrt{n}}. \quad (\text{A.45})$$



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

$$\sup_{|c_1| \leq M, c_2 > 0, s > 0} (c_2 \vee 1)^{-2} |V_2^{(n,p)}(c_1, c_2, s) - V_2^{(\infty,p)}(c_1, c_2, s)| \leq L' \cdot \frac{\log n}{\sqrt{n}}. \quad (\text{A.46})$$

For the term,

$$(c_2 \vee 1)^{-2} |V_2^{(\infty,p)}(c_1, c_2, s) - V_2^{(\infty,\infty)}(c_1, c_2, s)| \quad (\text{A.47})$$

$$= (c_2 \vee 1)^{-2} \left| \left( \mathbf{E}_{(\Lambda, W, G) \sim \mathcal{Q}_p} - \mathbf{E}_{(\Lambda, W, G) \sim \mathcal{Q}_\infty} \right) \tilde{f}_{c_1, c_2, s}(\Lambda, W, G) \right|, \quad (\text{A.48})$$

with

$$\tilde{f}_{c_1, c_2, s}(\Lambda, W, G) := \left( \frac{\Lambda^{-1/2} \mathbf{prox}_s \left( \Lambda^{1/2} \Pi_{W^\perp}(G) + \psi^{-1/2} [\partial_1 F_\kappa(c_1, c_2) - c_1 c_2^{-1} \partial_2 F_\kappa(c_1, c_2)] \Lambda^{1/2} W \right)}{\psi^{-1/2} c_2^{-1} \partial_2 F_\kappa(c_1, c_2)} \right)^2 \quad (\text{A.49})$$

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

$$\frac{(c_2 \vee 1)^{-2} |\tilde{f}_{c_1, c_2, s}(\Lambda, W, G)|}{1 + \Lambda^2 + W^2 + G^2} < \infty. \quad (\text{A.50})$$

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

$$\lim_{n \rightarrow \infty, p(n)/n = \psi} \sup_{|c_1| \leq M, c_2 > 0, s > 0} (c_2 \vee 1)^{-2} |V_2^{(\infty,p)}(c_1, c_2, s) - V_2^{(\infty,\infty)}(c_1, c_2, s)| = 0, \quad (\text{A.51})$$

$$\lim_{n \rightarrow \infty, p(n)/n = \psi} \sup_{|c_1| \leq M, c_2 > 0, s > 0} (c_2 \vee 1)^{-2} |V_2^{(n,p)}(c_1, c_2, s) - V_2^{(\infty,p)}(c_1, c_2, s)| = 0 \text{ a.s.} \quad (\text{A.52})$$

**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

$$\lim_{n \rightarrow \infty, p(n)/n = \psi} \hat{c}_1 = c_1^\star, \quad \lim_{n \rightarrow \infty, p(n)/n = \psi} \hat{c}_2 = c_2^\star, \quad \lim_{n \rightarrow \infty, p(n)/n = \psi} \hat{s} = s^\star \quad (\text{A.53})$$

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  $\delta$ , there exist  $n, p$  large enough, such that for  $(\hat{c}_1, \hat{c}_2, \hat{s})$

$$(1 \vee \hat{c}_2)^{-1} |V_1^{(n,p)}(\hat{c}_1, \hat{c}_2, \hat{s}) - V_1^{(\infty,\infty)}(\hat{c}_1, \hat{c}_2, \hat{s})| \leq \delta. \quad (\text{A.54})$$

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

$$(1 \vee \hat{c}_2)^{-1} |V_1^{(\infty, \infty)}(\hat{c}_1, \hat{c}_2, \hat{s})| \leq \delta, \quad (\text{A.55})$$

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

$$(\hat{c}_1, \hat{c}_2, \hat{s}) \in \mathcal{B}((c_1^*, c_2^*, s^*), \epsilon). \quad (\text{A.56})$$

Note this holds for arbitrary  $\epsilon$ . Therefore, we have proved Eqn. (A.53).

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

$$\|\hat{\theta}\|^2 \asymp \|\Lambda^{1/2} \hat{\theta}\|_2^2 = \langle w, \Lambda^{1/2} \hat{\theta} \rangle^2 + \|\Pi_{w^\perp} \hat{\theta}\|_2^2 = \hat{c}_1^2 + \hat{c}_2^2 \leq 2(c_1^*)^2 + 2(c_2^*)^2 < 4R^2 \quad (\text{A.57})$$

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

$$\hat{T}(\psi, \kappa; \hat{c}_1, \hat{c}_2, \hat{s}) := \psi^{-1/2} [\widehat{F}_\kappa(\hat{c}_1, \hat{c}_2) - \hat{c}_1 \partial_1 \widehat{F}_\kappa(\hat{c}_1, \hat{c}_2) - \hat{c}_2 \partial_2 \widehat{F}_\kappa(\hat{c}_1, \hat{c}_2)] - \hat{s} \quad (\text{A.58})$$

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  $c_1^*, c_2^*$ , say denoted by  $\mathcal{B}(c_1^*, R), \mathcal{B}(c_2^*, R)$ . It is easy to show that  $\widehat{F}_\kappa$  satisfies the uniform convergence bound

$$\lim_{n \rightarrow \infty, p(n)/n = \psi} \sup_{c_1, c_2 \in \mathcal{B}(c_1^*, R), \mathcal{B}(c_2^*, R)} |\widehat{F}_\kappa(c_1, c_2) - F_\kappa(c_1, c_2)| = 0 \quad a.s. \quad (\text{A.59})$$

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

$$\lim_{n \rightarrow \infty, p(n)/n = \psi} \hat{T}(\psi, \kappa; \hat{c}_1, \hat{c}_2, \hat{s}) \quad (\text{A.60})$$

$$= \lim_{n \rightarrow \infty, p(n)/n = \psi} \psi^{-1/2} [F_\kappa(\hat{c}_1, \hat{c}_2) - \hat{c}_1 \partial_1 F_\kappa(\hat{c}_1, \hat{c}_2) - \hat{c}_2 \partial_2 F_\kappa(\hat{c}_1, \hat{c}_2)] - \hat{s} \quad (\text{unif. conv.}) \quad (\text{A.61})$$

$$= \psi^{-1/2} [F_\kappa(c_1^*, c_2^*) - c_1^* \partial_1 F_\kappa(c_1^*, c_2^*) - c_2^* \partial_2 F_\kappa(c_1^*, c_2^*)] - s^* = T(\psi, \kappa). \quad (\text{A.62})$$

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

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

□

Below, we introduce a key lemma used in the uniform convergence proof in Proposition A.1. This 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  $1 - n^{-2}$ ,*

$$\sup_{|c_1| \leq M, c_2 > 0} |\partial_i \widehat{F}_\kappa(c_1, c_2) - \partial_i F_\kappa(c_1, c_2)| \leq C \cdot \frac{\log n}{\sqrt{n}}, \quad (\text{A.63})$$

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  $F_\kappa$ , 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

$$\partial_1 \widehat{F}_\kappa(c_1, c_2) = -\frac{\widehat{\mathbf{E}}_n[YZ_1 \sigma(\kappa - c_1 YZ_1 - c_2 Z_2)]}{(\widehat{\mathbf{E}}_n[\sigma^2(\kappa - c_1 YZ_1 - c_2 Z_2)])^{1/2}}, \quad (\text{A.64})$$

$$\partial_2 \widehat{F}_\kappa(c_1, c_2) = -\frac{\widehat{\mathbf{E}}_n[Z_2 \sigma(\kappa - c_1 YZ_1 - c_2 Z_2)]}{(\widehat{\mathbf{E}}_n[\sigma^2(\kappa - c_1 YZ_1 - c_2 Z_2)])^{1/2}}, \quad (\text{A.65})$$

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  $(M, \infty)$ .

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  $\epsilon$ -covering net as  $\mathcal{N}_\epsilon([-M, M] \times (0, M])$ , we know that on this bounded region, with probability at least  $1 - n^{-2}$ ,

$$\begin{aligned} & \sup_{(c_1, c_2) \in [-M, M] \times (0, M]} |\widehat{\mathbf{E}}_n[R_j(c_1, c_2)] - \mathbf{E}[R_j(c_1, c_2)]| \\ & \leq \sup_{(c'_1, c'_2) \in \mathcal{N}_\epsilon} |\widehat{\mathbf{E}}_n[R_j(c'_1, c'_2)] - \mathbf{E}[R_j(c'_1, c'_2)]| \\ & + \sup_{(c_1, c_2) \in [-M, M] \times (0, M]} \inf_{(c'_1, c'_2) \in \mathcal{N}_\epsilon} |\widehat{\mathbf{E}}_n[R_j(c_1, c_2)] - \widehat{\mathbf{E}}_n[R_j(c'_1, c'_2)]| \\ & + \sup_{(c_1, c_2) \in [-M, M] \times (0, M]} \inf_{(c'_1, c'_2) \in \mathcal{N}_\epsilon} |\mathbf{E}[R_j(c_1, c_2)] - \mathbf{E}[R_j(c'_1, c'_2)]| \\ & \lesssim \frac{\log \frac{1}{\epsilon^2}}{\sqrt{n}} + (\log n + 1)\epsilon \lesssim \frac{\log n}{\sqrt{n}}, \quad \forall j \in 0, 1, 2. \end{aligned} \quad (\text{A.66})$$

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

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

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

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

$$\partial_1 \widehat{F}_\kappa(c_1, c_2) = - \frac{\widehat{\mathbf{E}}_n[Y Z_1 \sigma(\kappa c_2^{-1} - c_1 c_2^{-1} Y Z_1 - Z_2)]}{(\widehat{\mathbf{E}}_n[\sigma^2(\kappa c_2^{-1} - c_1 c_2^{-1} Y Z_1 - Z_2)])^{1/2}}, \quad (\text{A.67})$$

$$\partial_2 \widehat{F}_\kappa(c_1, c_2) = - \frac{\widehat{\mathbf{E}}_n[Z_2 \sigma(\kappa c_2^{-1} - c_1 c_2^{-1} Y Z_1 - Z_2)]}{(\widehat{\mathbf{E}}_n[\sigma^2(\kappa c_2^{-1} - c_1 c_2^{-1} Y Z_1 - Z_2)])^{1/2}}. \quad (\text{A.68})$$

Now the regions for the parameters of interest are bounded since

$$(c_2^{-1}, c_1 c_2^{-1}) \in [0, 1/M] \times (-1, 1). \quad (\text{A.69})$$

Now define  $a = c_2^{-1}, b = c_1 c_2^{-1}$ ,  $\tilde{R}_1(a, b) := Y Z_1 \sigma(\kappa a - b Y Z_1 - Z_2)$ ,  $\tilde{R}_2(a, b) := Z_2 \sigma(\kappa a - b Y Z_1 - Z_2)$  and  $\tilde{R}_0(a, b) := \sigma^2(\kappa a - b Y Z_1 - Z_2)$  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  $\epsilon$ -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

$$\begin{aligned} & \sup_{(c_1, c_2) \in [-M, M] \times (M, \infty)} |\widehat{\mathbf{E}}_n[\tilde{R}_j(c_2^{-1}, c_1 c_2^{-1})] - \mathbf{E}[\tilde{R}_j(c_2^{-1}, c_1 c_2^{-1})]| \\ & \leq \sup_{(a, b) \in \mathcal{N}_\epsilon} |\widehat{\mathbf{E}}_n[\tilde{R}_j(a, b)] - \mathbf{E}[\tilde{R}_j(a, b)]| \\ & + \sup_{(c_1, c_2) \in [-M, M] \times (M, \infty)} \inf_{(a, b) \in \mathcal{N}_\epsilon} |\widehat{\mathbf{E}}_n[\tilde{R}_j(c_2^{-1}, c_1 c_2^{-1})] - \widehat{\mathbf{E}}_n[\tilde{R}_j(a, b)]| \\ & + \sup_{(c_1, c_2) \in [-M, M] \times (M, \infty)} \inf_{(a, b) \in \mathcal{N}_\epsilon} |\mathbf{E}[\tilde{R}_j(c_2^{-1}, c_1 c_2^{-1})] - \mathbf{E}[\tilde{R}_j(a, b)]| \\ & \lesssim \frac{\log \frac{1}{\epsilon^2}}{\sqrt{n}} + (\log n + 1)\epsilon \lesssim \frac{\log n}{\sqrt{n}}, \quad \forall j \in 0, 1, 2. \end{aligned} \quad (\text{A.70})$$

The proof can be completed following standard algebra based on the expression (A.67) and (A.68), since

$$\partial_j \widehat{F}_\kappa(c_1, c_2) = - \frac{\widehat{\mathbf{E}}_n[\tilde{R}_j(c_2^{-1}, c_1 c_2^{-1})]}{\sqrt{\widehat{\mathbf{E}}_n[\tilde{R}_0(c_2^{-1}, c_1 c_2^{-1})]}}. \quad (\text{A.71})$$

□

### A.3 Proof Outline for Generalization Error

*Proof of Theorem 3.2.* The proof follows by an adaptation of [76, Section E], on using Theorem 3.1 and Proposition A.1. Here we provide an outline of the argument. Note that since the model (2.1) involves Gaussian covariates, by rotation, we can equivalently express it as a model where all but the first coordinate of the true signal is zero. Thus,

$$\mathbb{P}_{(\mathbf{x}, \mathbf{y})}(\mathbf{y} \cdot \mathbf{x}^\top \hat{\theta}_{n, \ell_1} < 0) = \mathbb{P}\left(c_{1, n} Y Z_1 + \sqrt{1 - c_{1, n}^2} Z_2\right), \quad (\text{A.72})$$

where  $c_{1,n} = \frac{\langle \hat{\theta}_{n,\ell_1}, \theta_\star \rangle_\Lambda}{\|\hat{\theta}_{n,\ell_1}\|_\Lambda \|\theta_\star\|_\Lambda}$  and  $(Y, Z_1, Z_2)$  satisfies the joint distribution given by (2.7). Recall that  $\langle u, v \rangle_\Lambda = u^\top \Lambda v$ ,  $\|u\|_\Lambda = u^\top \Lambda u$ . Thus it suffices to show that

$$c_{1,n} \xrightarrow{\text{a.s.}} c_1^\star, \quad (\text{A.73})$$

where  $c_1^\star$  is defined following (3.4).

Now, recall that the min- $\ell_1$ -norm interpolant solves

$$\xi_{\psi, \kappa}^{(n,p)} = \min_{\|\theta\|_1 \leq \sqrt{p}} \frac{1}{\sqrt{p}} \|(\kappa \mathbf{1} - (y \odot X)\theta)_+\|_2.$$

For any compact set  $\Theta_p$ , define

$$\xi_{\psi, \kappa}^{(n,p)}(\Theta_p) = \min_{\theta \in \Theta_p} \frac{1}{\sqrt{p}} \|(\kappa \mathbf{1} - (y \odot X)\theta)_+\|_2.$$

If one can show that

$$\xi_{\psi, \kappa}^{(n,p)}(\Theta_p) > \xi_{\psi, \kappa}^{(n,p)}, \quad (\text{A.74})$$

then  $\mathbb{P}(\hat{\theta}_{n,\ell_1} \in \Theta_p) \rightarrow 0$  as  $n \rightarrow \infty$ . This is the key idea behind the proof. Naturally, this suggests choosing  $\Theta_p$  to be compact sets of the form

$$\Theta_p = [\theta : \|\theta\|_1 \leq \sqrt{p}, c_{1,n} \in [c_1^\star - \epsilon, c_1^\star + \epsilon]^c],$$

and establishing (A.74) for every  $\epsilon > 0$ . To formally establish this argument, define  $\hat{\kappa}_n = \min_{1 \leq i \leq n} y_i x_i^\top \hat{\theta}_{n,\ell_1}$  and note that

$$\hat{\kappa}_n = \sup_{\kappa} \{\xi_{\psi, \kappa}^{(n,p)} = 0\}. \quad (\text{A.75})$$

Further, define  $\xi_{\psi, \kappa}^{(n,p)}(c) = \min_{\|\theta\|_1 \leq \sqrt{p}, \frac{\langle \theta, \theta_\star \rangle_\Lambda}{\|\theta\|_\Lambda \|\theta_\star\|_\Lambda} = c} \frac{1}{\sqrt{p}} \|(\kappa \mathbf{1} - (y \odot X)\theta)_+\|_2$  and note that

$$c_{1,n} \in \{c : \xi_{\psi, \hat{\kappa}_n}^{(n,p)}(c) = 0\}. \quad (\text{A.76})$$

Now, for any  $c_1, c_2 \in [-1, 1]$ , define

$$\xi_{\psi, \kappa}^{(n,p)}(c_1, c_2) = \min \left\{ \min_{c \leq c_1} \xi_{\psi, \kappa}^{(n,p)}(c), \min_{c \geq c_2} \xi_{\psi, \kappa}^{(n,p)}(c) \right\}.$$

To show (A.73), from (A.75)-(A.76), the final step then involves establishing that for any  $\epsilon > 0$

$$\lim_{n, p \rightarrow \infty, p/n \rightarrow \psi} \mathbb{P}[\xi_{\psi, \hat{\kappa}_n}^{(n,p)}(c_1^\star - \epsilon, c_1^\star + \epsilon) > 0] = 1.$$

This can be established by analytic arguments similar to [76, Section E], on using the limiting characterizations of  $\hat{\kappa}_n$  and  $\xi_{\psi, \hat{\kappa}_n}^{(n,p)}$  from Theorem 3.1 and Proposition A.1. □

#### A.4 Uniqueness Results

We next present the proof of Proposition 3.1. The first part of the proof is the same as that in [76], but we include it for the sake of completeness. The second part of the proof, though draws inspiration from [76], has different executions in this  $\ell_1$  case.

*Proof of Proposition 3.1.* To analyze the equation system (3.9), we will, in fact, begin by examining the objective function in (A.9) as a function of  $h$ , that is, define

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

and consider the optimization problem

$$\text{minimize} \quad \mathcal{R}_{\psi, \kappa, Q_\infty}(h) \quad \text{s.t.} \quad \|h\|_{L_1(Q_\infty)} \leq 1. \quad (\text{A.77})$$

Due to the form of the constraint set, it follows from the Banach-Alaoglu theorem that the minimum is achieved for some  $h^\star \in L_2(Q_\infty)$ . Further, using the fact that  $F_\kappa$  is convex and increasing with respect to the second argument (see [76, Lemma 5.3.(b)]), it can be shown that the function  $h \rightarrow \mathcal{R}_{\psi, \kappa, Q_\infty}$  is strictly convex. This immediately implies uniqueness of the minimizer, in the sense that, given two minimizers  $h^\star$  and  $\tilde{h}$ , one must have  $\mathbb{P}[h^\star \neq \tilde{h}] = 0$ . Then the unique minimizer is determined by the KKT conditions, which in this case can be expressed as

$$\begin{aligned} \Lambda^{1/2} \Pi_{W^\perp}(G) + \psi^{-1/2} \Lambda^{1/2} [\partial_1 F_\kappa(c_1, c_2) W + \partial_2 F_\kappa(c_1, c_2) \Pi_{W^\perp}(Z)] + s \cdot \partial \|h\|_{L_1(Q_\infty)} &= 0, \\ s(1 - \|h\|_{L_1(Q_\infty)}) &= 0, \\ s \geq 0, \|h\|_{L_1(Q_\infty)} &\leq 1. \end{aligned} \quad (\text{A.78})$$

Above,  $Z$  is given by

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

and

$$c_1 = \langle W, \Lambda^{1/2} h \rangle_{L_2(Q_\infty)}, c_2 = \|\Pi_{W^\perp}(\Lambda^{1/2} h)\|_{L_2(Q_\infty)}. \quad (\text{A.79})$$

We claim that any solution of (A.78) and the associated dual variable  $s$  satisfy  $s > 0$  and  $\|\Pi_{W^\perp}(\Lambda^{1/2} h)\| > 0$ . The former follows directly from [76, Section B.3.3], but we describe the detail here for the sake of completion. Recall that  $(\Lambda, W) \sim \mu$  defined in (2.4). From properties of  $\mu$  it follows that  $\Lambda > 0, \|W\| = 1$ . Suppose if possible that  $s = 0$ , then (A.78) implies that

$$\Pi_{W^\perp}(G) + \psi^{-1/2} [\partial_1 F_\kappa(c_1, c_2) W + \partial_2 F_\kappa(c_1, c_2) \Pi_{W^\perp}(Z)] = 0. \quad (\text{A.80})$$

Taking inner products with  $W$  on both sides, we obtain  $\psi^{-1/2} [\partial_1 F_\kappa(c_1, c_2)] = 0$ . Using this relation back in (A.80), we obtain that

$$\begin{aligned} \psi^{-1/2} \partial_2 F_\kappa(c_1, c_2) \Pi_{W^\perp}(Z) &= -\Pi_{W^\perp}(G) \\ \implies \psi^{-1/2} \partial_2 F_\kappa(c_1, c_2) &\geq \|\Pi_{W^\perp}(G)\|, \end{aligned}$$



by taking norm on both sides and noting that (i) the partial derivative with respect to the second coordinate of  $F_\kappa(\cdot, \cdot)$  is always positive (ii)  $\|\Pi_{W^\perp}(Z)\| \leq \|Z\| \leq 1$ . From [76, Lemma B.1], we know that if  $(c_1, c_2)$  is a tuple satisfying  $\partial_1 F_\kappa(c_1, c_2) = 0$ , then the partial derivative with respect to the second coordinate at the same point can be at most square root of the separability threshold, that is,  $\partial_2 F_\kappa(c_1, c_2) \leq \min_{c \in \mathbb{R}} F_0(c, 1) = \sqrt{\psi^\star(\rho, f)}$ . Together this yields that  $\sqrt{\psi} \|\Pi_{W^\perp}(G)\| \leq \sqrt{\psi^\star(\rho, f)}$ , which, from the definition of  $\psi^\downarrow(\kappa)$  (3.12), and the fact that  $W, G$  are independent, contradicts our assumption that  $\psi > \psi^\downarrow(\kappa)$  in the hypothesis of the proposition.

We next proceed to show that for any solution  $h$ ,  $c_2 = \|\Pi_{W^\perp}(\Lambda^{1/2}h)\| > 0$ . Suppose by contradiction that  $c_2 = 0$ . By decomposing  $h$  in the direction of  $W$  and  $W^\perp$ , observe that in this case

$$\Lambda^{1/2}h = c_1 W, \quad (\text{A.81})$$

where recall the definition of  $c_1$  from (A.79). Since we established  $s > 0$ , for any solution,  $\|h\|_{L_1(\mathcal{Q}_\infty)} = 1$ . This yields that in this case,

$$|c_1| = \zeta \quad (\text{A.82})$$

Now divide the problem into two cases: Case (i):  $c_1 > 0$  and Case (ii):  $c_1 < 0$ . Here we only show the argument for Case (i), since the other case follows similarly. From the first equation in (A.78), we have

$$\Pi_{W^\perp}(G) + \psi^{-1/2} [\partial_1 F_\kappa(c_1, c_2)W + \partial_2 F_\kappa(c_1, c_2)\Pi_{W^\perp}(Z)] + s \cdot \Lambda^{-1/2} \partial \|h\|_{L_1(\mathcal{Q}_\infty)} = 0 \quad (\text{A.83})$$

Taking inner product with  $W$  and using  $\mathbb{E}[W^2] = 1$ , and using the facts that  $c_1 = \zeta, c_2 = 0$ , (A.81), and  $\|h\|_{L_1(\mathcal{Q}_\infty)} = 1$  for a solution  $h$ , we obtain

$$\begin{aligned} \psi^{-1/2} \partial_1 F_\kappa(\zeta, 0) + s \langle \Lambda^{-1/2} W, \partial \|h\|_{L_1(\mathcal{Q}_\infty)} \rangle &= 0 \\ \psi^{-1/2} \partial_1 F_\kappa(\zeta, 0) + s c_1^{-1} \langle h, \partial \|h\|_{L_1(\mathcal{Q}_\infty)} \rangle &= 0 \\ \psi^{-1/2} \partial_1 F_\kappa(\zeta, 0) + s c_1^{-1} \|h\|_{L_1(\mathcal{Q}_\infty)} &= 0 \\ s = -c_1 \psi^{-1/2} \partial_1 F_\kappa(\zeta, 0) & \end{aligned} \quad (\text{A.84})$$

Since  $s > 0$ , this yields that  $\partial_1 F_\kappa(\zeta, 0) < 0$ , which implies that by definition,  $\psi$  should be above the threshold  $\psi_+(\kappa)$  that satisfies

$$\mathbb{E} \left[ \left\{ \psi_+(\kappa)^{1/2} \Pi_{W^\perp}(G) + \partial_1 F_\kappa(\zeta, 0) (W - \zeta \Lambda^{-1/2} \text{sign}(\zeta \Lambda^{-1/2} W)) \right\}^2 \right] = \partial_2^2 F_\kappa(\zeta, 0). \quad (\text{A.85})$$

Now plugging (A.84) back in (A.83) and using (A.81), we obtain that

$$\begin{aligned} \psi^{1/2} \Pi_{W^\perp}(G) + [\partial_1 F_\kappa(\zeta, 0)W + \partial_2 F_\kappa(\zeta, 0)\Pi_{W^\perp}(Z)] - c_1 \partial_1 F_\kappa(\zeta, 0) \cdot \Lambda^{-1/2} \text{sign}(\zeta \Lambda^{-1/2} W) &= 0 \\ \psi^{1/2} \Pi_{W^\perp}(G) + \partial_1 F_\kappa(\zeta, 0)(W - c_1 \Lambda^{-1/2} \text{sign}(\zeta \Lambda^{-1/2} W)) &= -\partial_2 F_\kappa(\zeta, 0) \Pi_{W^\perp}(Z) \end{aligned} \quad (\text{A.86})$$

If we take  $\ell_2$  norm on both sides of the above, and recall that  $\|\Pi_{W^\perp}(Z)\|_2 < 1$  for  $c_2 = 0$ , we obtain

$$\mathbb{E}[\{\psi^{1/2}\Pi_{W^\perp}(G) + \partial_1 F_\kappa(\zeta, 0)(W - c_1 \Lambda^{-1/2} \text{sign}(\zeta \Lambda^{-1/2} W))\}^2] < \partial_2 F_\kappa(\zeta, 0)^2.$$

However, this contradicts the range of  $\psi$  determined by (A.85). The case of  $c_1 < 0$  can be similarly handled on recalling the fact that, by definition,  $\psi > \psi_-(\kappa)$  introduced in (3.12). Thus, we conclude that  $c_2 > 0$ .

Now that we have established that  $c_2$  and  $s$  must be strictly positive when  $(c_1, c_2, s)$  solves our equation system, we can proceed to explicitly identify the formula for the solution  $h^\star$  to (A.78). The KKT conditions yield that

$$\begin{aligned} \psi^{-1/2} c_2^{-1} \partial_2 F_\kappa(c_1, c_2) \Lambda^{1/2} h + s \cdot \Lambda^{-1/2} \partial \|h\|_{L_1(\mathcal{Q}_\infty)} \\ = -(\Pi_{W^\perp}(G) + \psi^{-1/2} [\partial_1 F_\kappa(c_1, c_2) - c_1 c_2^{-1} \partial_2 F_\kappa(c_1, c_2)] W) \end{aligned} \quad (\text{A.87})$$

From [76, Section B.3],  $\partial_2 F_\kappa(\zeta, 0) > 0$ , so we may rewrite the above as follows:

$$h + \frac{s \cdot \Lambda^{-1/2} \partial \|h\|_{L_1(\mathcal{Q}_\infty)}}{\psi^{-1/2} c_2^{-1} \partial_2 F_\kappa(c_1, c_2) \Lambda^{1/2}} = -\frac{(\Pi_{W^\perp}(G) + \psi^{-1/2} [\partial_1 F_\kappa(c_1, c_2) - c_1 c_2^{-1} \partial_2 F_\kappa(c_1, c_2)] W)}{\psi^{-1/2} c_2^{-1} \partial_2 F_\kappa(c_1, c_2) \Lambda^{1/2}}.$$

Now the above implies that the solution  $h^\star$  is given by

$$\begin{aligned} h^\star &= \text{prox}_{\frac{s \cdot \Lambda^{-1/2}}{\psi^{-1/2} c_2^{-1} \partial_2 F_\kappa(c_1, c_2) \Lambda^{1/2}}} \left( -\frac{(\Pi_{W^\perp}(G) + \psi^{-1/2} [\partial_1 F_\kappa(c_1, c_2) - c_1 c_2^{-1} \partial_2 F_\kappa(c_1, c_2)] W)}{\psi^{-1/2} c_2^{-1} \partial_2 F_\kappa(c_1, c_2) \Lambda^{1/2}} \right), \\ &= -\frac{\Lambda^{-1}}{\psi^{-1/2} c_2^{-1} \partial_2 F_\kappa(c_1, c_2)} \text{prox}_s(\Lambda^{1/2} G + \psi^{-1/2} [\partial_1 F_\kappa(c_1, c_2) - c_1 c_2^{-1} \partial_2 F_\kappa(c_1, c_2)] \Lambda^{1/2} W) \end{aligned}$$

Plugging this in the system

$$c_1 = \langle \Lambda^{1/2} h^\star, W \rangle_{L_2(\mathcal{Q}_\infty)}, \quad c_1^2 + c_2^2 = \|\Lambda^{1/2} h^\star\|_{L_2(\mathcal{Q}_\infty)}^2, \quad \|h^\star\|_{L_1(\mathcal{Q}_\infty)} = 1 \quad (\text{A.88})$$

yields the fixed point equations (3.9). 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.87) are also unique and, furthermore,  $c_2$  and  $s$  are strictly positive.  $\square$

We obtain a key representation for  $\tilde{\xi}_{\psi, \kappa}^{(\infty, \infty)}(\Lambda, W, G)$  as a byproduct of the above. On taking inner products with  $\Lambda^{1/2} h$  on both sides of (A.87) leads to the following.

**Corollary A.1.** *Under the assumptions of Proposition 3.1, the minimum value of the optimization problem (A.77) is given by*

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

where  $(c_1, c_2, s) \in \mathbb{R} \times \mathbb{R}_{>0} \times \mathbb{R}_{>0}$  forms the unique solution to (3.9). 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  $\ell_1$  geometry. Therefore, the arguments in Section A.2 naturally extend to the  $\ell_q$  geometry (as long as  $q \leq 2$ ), and those in the current section can also be extended to show uniqueness of the system (3.21) on changing the definition of  $\zeta$  and establishing bounds on  $\langle W, \Lambda^{-1/2} \partial \|h\|_{L_q(\mathcal{Q}_\infty)} \rangle$  appropriately.

## 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* from Section 2, with bounded continuous  $|X_{ij}| \leq M$  and a shrinkage on the learning rate (the specifics will be made clear in the proof below). Note that in the discrete case (Case (1)), Steps (a) and (b) in the *Boosting Algorithm* from Section 2 could be replaced by

$$v_{t+1} := \arg \min_{v \in \{\pm e_j\}_{j \in [p]}} \sum_{i \in [n]} \eta_t[i] \cdot \mathbb{I}_{y_i x_i^\top v \leq 0}$$

$$\alpha_t = \frac{1}{2} \log \left( \frac{1 - \sum_{i \in [n]} \eta_t[i] \cdot \mathbb{I}_{y_i x_i^\top v_{t+1} \leq 0}}{\sum_{i \in [n]} \eta_t[i] \cdot \mathbb{I}_{y_i x_i^\top v_{t+1} \leq 0}} \right).$$

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

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

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

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

One can verify that  $R$  is 1-strongly convex w.r.t. the  $\ell_1$  norm, and that  $R^\star$  is 1-strongly smooth w.r.t. the  $L_\infty$  norm.

First, let us recall the dual formulation of  $\ell_1$ -margin, and the von Neumann's minimax theorem

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

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

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

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

$$\min_{v \in \{\pm e_j\}_{j \in [p]}} \sum_{i \in [n]} \eta_t[i] \cdot \mathbb{I}_{y_i x_i^\top v \leq 0} = \min_{v \in \{\pm e_j\}_{j \in [p]}} \sum_{i \in [n]} \eta_t[i] \cdot \mathbb{I}_{-y_i x_i^\top v \geq 0} \quad (\text{A.92})$$

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

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

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

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

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

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

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

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

For the (2) *Boosting Algorithm* from Section 2, with  $|X_{ij}| \leq M$ , define a shrinkage on the learning rate  $\alpha_t(\beta)$  with a constant factor  $\beta > 0$ ,

$$\alpha_t(\beta) = \beta \cdot \eta_t^\top Zv_{t+1} . \quad (\text{A.97})$$

A good choice of  $\beta$  will be clear in a second. Then

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

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

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

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

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

The proof is now complete. □

*Proof of Corollary 5.1.* The proof follows from Proposition 5.1 and a re-scaling technique in [108]'s asymptotic analysis. Here instead, we spell out a non-asymptotic result. For any  $\kappa > 0$

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

$$\leq \exp(\kappa \|\theta_t\|_1) \exp(R^*(-Z\theta_t)) , \quad (\text{A.101})$$

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

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

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

In addition, due to the coordinate update of  $\theta_t$ , we know

$$\|\theta_T\|_1 \leq \sum_{t=0}^{T-1} \|\alpha_t v_{t+1}\|_1 \leq \sum_{t=0}^{T-1} \beta\gamma_t . \quad (\text{A.104})$$

Therefore

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

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

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

With the choice of

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

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

we know that

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

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  $\kappa = \kappa_{n,\ell_1} \cdot (1 - \epsilon)$

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

we must have that

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

□

*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

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

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

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

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

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

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

$$\gamma_t = \|Z^\top \eta_t\|_{q_\star} \geq \min_{\eta \in \Delta} \max_{\|\theta\|_q \leq 1} \eta^\top Z \theta = \max_{\|\theta\|_q \leq 1} \min_{i \in [n]} y_i x_i^\top \theta = \kappa_{n, \ell_q}. \quad (\text{A.116})$$

The proof is complete if we plug in

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

□

For completeness, we show that the min- $\ell_1$ -norm interpolation, is equivalent to the max- $\ell_1$ -margin formulation. We use this fact several places in the main text.

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

$$\text{Formulation I: } I^\star := \max \left\{ \kappa \mid \exists \theta, \|\theta\|_1 \leq 1, \text{ s.t. } \forall i \leq n, y_i x_i^\top \theta \geq \kappa \right\} \quad (\text{A.118})$$



$$\text{Formulation II: } II^\star := \min \|\theta\|_1, \text{ s.t. } \forall i \leq n, y_i x_i^\top \theta \geq 1 \quad (\text{A.119})$$

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

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

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

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

□

## B Extended Derivations

We collect here the detailed derivations in Section 3.5, where robustness of the assumptions is investigated.

### B.1 Derivations in Section 3.5.1

This section provides details on the results mentioned in Section 3.5.1. Recall the generalization of GMMs considered in (3.28): we observe i.i.d. samples  $(x_i, y_i)$  such that  $\mathbb{P}[y_i = 1] = v = 1 - \mathbb{P}[y_i = -1]$  and  $x_i = y_i \theta_\star + m_i \tilde{\theta} + \tilde{x}_i$ . Here,  $\tilde{x}_i \sim \mathcal{N}(0, \Lambda)$  with  $\Lambda$  diagonal,  $(y_i, m_i, \tilde{x}_i)$  are independent, and  $m_i$  is symmetric around zero so that  $y_i \odot m_i \stackrel{d}{=} m_i$ . Stacking  $\tilde{x}_i$ 's as the rows within a matrix  $\tilde{X}$ , we observe that  $\tilde{X} = Z\Lambda^{1/2}$ , where  $Z$  has i.i.d.  $\mathcal{N}(0, 1)$  entries. Similarly, stacking  $x_i$ 's into rows of  $X$ , we obtain that

$$X = y\theta_\star^\top + m\tilde{\theta}^\top + \tilde{X}, \quad (\text{B.1})$$

where  $y = [y_1, \dots, y_n]^\top, m = [m_1, \dots, m_n]^\top$ . Recall from (5.2) that the max-min- $\ell_1$ -margin properties can be characterized by analyzing the following optimization problem:

$$\xi_n = \min_{\|\theta\|_1 \leq \sqrt{p}} \max_{\|\lambda\|_2 \leq 1, \lambda \geq 0} \frac{1}{\sqrt{p}} [\lambda^\top (\kappa \mathbf{1} - (y \odot X)\theta)].$$

In the context of our model (B.1),  $y \odot X \stackrel{d}{=} \mathbf{1}\theta_\star^\top + m\tilde{\theta}^\top + Z\Lambda^{1/2}$ , therefore  $\xi_n$  simplifies to

$$\xi_n = \min_{\|\theta\|_1 \leq \sqrt{p}} \max_{\|\lambda\|_2 \leq 1, \lambda \geq 0} \frac{1}{\sqrt{p}} [\lambda^\top \{(\kappa - \langle \theta_\star, \theta \rangle) \mathbf{1} - \langle \tilde{\theta}, \theta \rangle m - Z\Lambda^{1/2} \theta\}],$$

and by an application of CGMT, this is asymptotically equivalent to analyzing the following optimization problem

$$\min_{\|\theta\|_1 \leq \sqrt{p}} \max_{\|\lambda\|_2 \leq 1, \lambda \geq 0} \frac{1}{\sqrt{p}} [\langle \lambda, (\kappa - \langle \theta_\star, \theta \rangle) \mathbf{1} - \langle \tilde{\theta}, \theta \rangle m - \|\Lambda^{1/2} \theta\|_2 z \rangle - \|\lambda\|_2 \langle g, \Lambda^{1/2} \theta \rangle],$$

where  $z, g$  are independent vectors with entries i.i.d.  $\mathcal{N}(0, 1)$ . Maximizing over  $\lambda$ , this further reduces to

$$\min_{\|\theta\|_1 \leq \sqrt{p}} \left[ \frac{1}{\sqrt{p}} \langle g, \Lambda^{1/2} \theta \rangle + \|\nu_+\|_2 \right], \quad (\text{B.2})$$

where  $\nu = (\kappa - \langle \theta_\star, \theta \rangle) \mathbf{1} - \langle \tilde{\theta}, \theta \rangle m - \|\Lambda^{1/2} \theta\|_2 z$ . Define

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

where  $M, \tilde{Z}$  denote random vectors with distribution  $M_i \stackrel{d}{=} m_i$  and  $\tilde{Z}_i \sim \mathcal{N}(0, 1)$ , all entries i.i.d. with  $M, Z$  independent of each other and  $\hat{\mathbb{E}}_n$  denoting the corresponding empirical distribution. With this notation, (B.2) simplifies to

$$\min_{\|\theta\|_1 \leq \sqrt{p}} \left[ \frac{1}{p} \langle g, \Lambda^{1/2} \theta \rangle + \psi^{-1/2} \hat{F}_\kappa(\langle \theta_\star, \theta \rangle, \langle \tilde{\theta}, \theta \rangle, \|\Lambda^{1/2} \theta\|_2) \right]. \quad (\text{B.3})$$

Using tricks similar to those in Proposition (A.1), we have that  $\xi_n$  must converge to the following infinite-dimensional version

$$\xi_\infty := \min_{\|h\|_{L_1(Q)} \leq 1} [\langle G, \Lambda^{1/2} h \rangle_{L_2(Q)} + \psi^{-1/2} F_\kappa(\langle h_\star, h \rangle, \langle \tilde{h}, h \rangle, \|\Lambda^{1/2} h\|_{L_2(Q)})], \quad (\text{B.4})$$

where  $h_\star, \tilde{h}$  correspond to  $\theta_\star, \tilde{\theta}$  respectively. Here we use the same trick as in (5.10) and go over to the space  $\{h : \mathbb{R}^4 \rightarrow \mathbb{R}, h \in \mathcal{L}^2(Q)\}$ , where  $Q = \mu \otimes \mathcal{N}(0, 1)$  with  $\mu$  given as follows: the empirical probability distributions  $\sum_{j=1}^p \delta_{(\lambda_i, \sqrt{p} \theta_\star^\top e_i, \sqrt{p} \tilde{\theta}^\top e_i)} / p \xrightarrow{W_2} \mu$ . To rigorize these arguments, we assume that the data is in the asymptotically linearly separable regime. Note that the exact threshold for separability here will be different from  $\psi^\star$  since the data-generating scheme is different in this context. With  $F_\kappa$  denoting  $F_\kappa(c_1, c_2, c_3)$ , the limiting version of  $\hat{F}_\kappa$ , where  $c_1 = \langle h_\star, h \rangle, c_2 = \langle \tilde{h}, h \rangle, c_3 = \|\Lambda^{1/2} h\|_{L_2(Q)}$ , the KKT conditions corresponding to  $\xi_\infty$  can then be characterized as follows,

$$\Lambda^{1/2} G + \psi^{-1/2} [\partial_1 F_\kappa h_\star + \partial_2 F_\kappa \tilde{h} + \Lambda^{1/2} \partial_3 F_\kappa Z] + s \partial \|h\|_{L_1(Q)} = 0,$$

where  $Z = \Lambda^{1/2} h / \|\Lambda^{1/2} h\|$  if the denominator is strictly positive, and  $Z'$  with  $\|Z'\| \leq 1$  when the denominator is zero. Rewriting things, we obtain

$$\Lambda^{1/2} G + \psi^{-1/2} [\partial_1 F_\kappa h_\star + \partial_2 F_\kappa \tilde{h}] + \psi^{-1/2} \Lambda c_3^{-1} \partial_3 F_\kappa h + s \partial \|h\|_{L_1(Q)} = 0.$$

Using properties of the proximal mapping operator, this yields

$$h_{\text{sol}} = - \frac{\text{prox}_s(\Lambda^{1/2} G + \psi^{-1/2} (\partial_1 F_\kappa h_\star + \partial_2 F_\kappa \tilde{h}))}{\Lambda \psi^{-1/2} c_3^{-1} \partial_3 F_\kappa}. \quad (\text{B.5})$$

Assuming that  $(\Lambda, h_\star, \tilde{h}, G) \sim Q = \mu \otimes \mathcal{N}(0, 1)$ , the system of equations governing the behavior of

the max- $\ell_1$ -margin and min- $\ell_1$ -interpolant is then given by

$$\begin{aligned} c_1 &= \mathbb{E}_{(\Lambda, h_\star, \tilde{h}, G) \sim Q} h_\star h_{\text{sol}} \\ c_2 &= \mathbb{E}_{(\Lambda, h_\star, \tilde{h}, G) \sim Q} \tilde{h} h_{\text{sol}} \\ c_3^2 &= \mathbb{E}_{(\Lambda, h_\star, \tilde{h}, G) \sim Q} (\Lambda^{1/2} h_{\text{sol}})^2 \\ 1 &= \mathbb{E}_{(\Lambda, h_\star, \tilde{h}, G) \sim Q} |h_{\text{sol}}| \end{aligned} \quad (\text{B.6})$$

For the formal argument that  $\xi_n \rightarrow \xi_\infty$  as  $n, p \rightarrow \infty$ , we assume that the aforementioned equation system admits a unique solution. We expect that arguments similar to Proposition (3.1) can be used to prove this in the regime where the data is asymptotically linearly separable.

The aforementioned arguments for the model in (B.1) naturally extend to the following,

$$x_i = y_i \theta_\star + \sum_{c=1}^{\ell} m_{i,c} \tilde{\theta}_c + \tilde{x}_i, \quad (\text{B.7})$$

where  $\tilde{x}_i \sim \mathcal{N}(0, \Lambda)$ ,  $\Lambda$  diagonal,  $y \odot (m_{i,1}, \dots, m_{i,\ell}) \stackrel{d}{=} (m_{i,1}, \dots, m_{i,\ell})$  for all  $i$  and  $y_i$ 's,  $m_i$ 's,  $\tilde{x}_i$ 's are independent. Note that, in this data generation scheme, the covariance between features is a rank  $\ell$  perturbation of a diagonal. Define  $\mu_\ell$  to be the probability distribution given by the limit  $\sum_{j=1}^p \delta_{(\lambda_i, \sqrt{p}\theta_\star^\top e_i, \sqrt{p}\tilde{\theta}_1^\top e_i, \sqrt{p}\tilde{\theta}_2^\top e_i, \dots, \sqrt{p}\tilde{\theta}_\ell^\top e_i)} \xrightarrow{W_2} \mu_\ell$ , and let  $(\Lambda, h_\star, \tilde{h}_1, \tilde{h}_2, \dots, \tilde{h}_\ell, G) \sim Q_\ell = \mu_\ell \otimes \mathcal{N}(0, 1)$ . Then the system of equations governing the behavior of the max- $\ell_1$ -margin and min- $\ell_1$ -interpolant is given by

$$\begin{aligned} c_1 &= \mathbb{E}_{(\Lambda, h_\star, \tilde{h}_1, \dots, \tilde{h}_\ell, G) \sim Q_\ell} h_\star h_{\text{sol}, \ell} \\ c_{i+1} &= \mathbb{E}_{(\Lambda, h_\star, \tilde{h}_1, \dots, \tilde{h}_\ell, G) \sim Q_\ell} \tilde{h}_i h_{\text{sol}, \ell}, \quad i = 1, \dots, \ell, \\ c_{\ell+2}^2 &= \mathbb{E}_{(\Lambda, h_\star, \tilde{h}_1, \dots, \tilde{h}_\ell, G) \sim Q_\ell} (\Lambda^{1/2} h_{\text{sol}, \ell})^2 \\ 1 &= \mathbb{E}_{(\Lambda, h_\star, \tilde{h}_1, \dots, \tilde{h}_\ell, G) \sim Q_\ell} |h_{\text{sol}}|, \end{aligned}$$

where  $h_{\text{sol}, \ell}$  is defined as follows

$$h_{\text{sol}, \ell} := -\frac{\text{prox}_s(\Lambda^{1/2} G + \psi^{-1/2}(\partial_1 F_\kappa h_\star + \partial_2 F_\kappa \tilde{h}_1 + \partial_3 F_\kappa \tilde{h}_2 + \dots + \partial_{\ell+1} F_\kappa \tilde{h}_\ell))}{\Lambda \psi^{-1/2} c_{\ell+2}^{-1} \partial_{\ell+2} F_\kappa}$$

and  $F_\kappa$  equals

$$F_\kappa(c_1, \dots, c_{\ell+2}) = \sqrt{\mathbb{E}(\kappa - c_1 - c_2 M_1 - c_3 M_2 - \dots - c_{\ell+1} M_\ell - c_{\ell+2} \tilde{Z})_+^2};$$

here  $\tilde{Z} \sim \mathcal{N}(0, 1)$ , independently of  $(M_1, \dots, M_\ell)$ , which has the same distribution as  $(m_{i,1}, \dots, m_{i,\ell})$ .

## B.2 Derivations in Section 3.5.2

*Proof of Theorem 3.4.* To overcome the difficulty of non-differentiability (due to  $\ell_1$ ) and the non-strongly convexity of our problem, we need to introduce a Gaussian smoothing technique and an extra  $\ell_2$  regularization term. To start, define the ReLU function

$$h(t) = \max(t, 0) \quad (\text{B.8})$$

and the smoothed version of the ReLU

$$h_\delta(t) := \mathbf{E}_{\mathbf{g} \sim \mathcal{N}(0,1)} [h(t + \delta \mathbf{g})] \quad (\text{B.9})$$

For the purposes of this section, we denote  $x_i$  (resp.  $\tilde{x}_i$ ) to mean the random features  $a_i$  (resp.  $b_i$ ), where  $a_i, b_i$  are as defined in Section 3.5.2.

To prove (3.32), we define for fixed  $\kappa, \lambda > 0$ , the following perturbed Lagrangian that is strongly convex in  $\theta$

$$\begin{aligned} \mathcal{L}_k^{\kappa, \lambda}(\theta ; \epsilon, \delta) &:= \\ &\sum_{i=1}^k h_\delta^2(\kappa - y_i \frac{1}{\sqrt{p}} x_i^\top \theta) + \sum_{i=k+1}^n h_\delta^2(\kappa - y_i \frac{1}{\sqrt{p}} \tilde{x}_i^\top \theta) + \lambda \sum_{j=1}^p (h_\delta(\theta_j) + h_\delta(-\theta_j) - 1) + \epsilon \sum_{j=1}^p \frac{1}{2} \theta_j^2 \\ \mathcal{L}_{\setminus k}^{\kappa, \lambda}(\theta ; \epsilon, \delta) &:= \\ &\sum_{i=1}^{k-1} h_\delta^2(\kappa - y_i \frac{1}{\sqrt{p}} x_i^\top \theta) + \sum_{i=k+1}^n h_\delta^2(\kappa - y_i \frac{1}{\sqrt{p}} \tilde{x}_i^\top \theta) + \lambda \sum_{j=1}^p (h_\delta(\theta_j) + h_\delta(-\theta_j) - 1) + \epsilon \sum_{j=1}^p \frac{1}{2} \theta_j^2. \end{aligned} \quad (\text{B.10})$$

Further, define

$$\begin{aligned} \Phi_k^{\kappa, \lambda}(\epsilon, \delta) &:= \min_{\theta \in \mathbb{R}^p} \mathcal{L}_k^{\kappa, \lambda}(\theta ; \epsilon, \delta) \\ \Phi_{\setminus k}^{\kappa, \lambda}(\epsilon, \delta) &:= \min_{\theta \in \mathbb{R}^p} \mathcal{L}_{\setminus k}^{\kappa, \lambda}(\theta ; \epsilon, \delta) \\ \theta_{\setminus k}^\star &:= \arg \min_{\theta \in \mathbb{R}^p} \mathcal{L}_{\setminus k}^{\kappa, \lambda}(\theta ; \epsilon, \delta) \end{aligned} \quad (\text{B.11})$$

and finally the leave-one-out Hessian

$$H_{\setminus k}(\epsilon) := \nabla_\theta^2 \mathcal{L}_{\setminus k}(\theta ; \lambda ; \epsilon, \delta) |_{\theta = \theta_{\setminus k}^\star} \succeq \epsilon \mathbf{I}_p \quad (\text{B.12})$$

with the full expression

$$\begin{aligned} H_{\setminus k}(\epsilon) &= \frac{2}{p} \sum_{i=1}^{k-1} h_\delta(\kappa - y_i \frac{1}{\sqrt{p}} x_i^\top \theta_{\setminus k}^\star) h'_\delta(\kappa - y_i \frac{1}{\sqrt{p}} x_i^\top \theta_{\setminus k}^\star) \cdot x_i x_i^\top \\ &\quad + \frac{2}{p} \sum_{i=k+1}^n h_\delta(\kappa - y_i \frac{1}{\sqrt{p}} \tilde{x}_i^\top \theta_{\setminus k}^\star) h'_\delta(\kappa - y_i \frac{1}{\sqrt{p}} \tilde{x}_i^\top \theta_{\setminus k}^\star) \cdot \tilde{x}_i \tilde{x}_i^\top \\ &\quad + \lambda \text{diag} \{ h''_\delta(\theta_{\setminus k, j}^\star) + h''_\delta(-\theta_{\setminus k, j}^\star) \} + \epsilon \mathbf{I}_p. \end{aligned} \quad (\text{B.13})$$

Our goal is to show  $\Phi_n(A, \lambda) - \Phi_n(B, \lambda) \xrightarrow{\mathbb{P}} 0$ , where these are defined in Section 3.5.2. In our notation here, this reduces to establishing that for all  $\lambda > 0$ ,

$$\frac{1}{p} \Phi_n^{\kappa, \lambda}(0, 0) - \frac{1}{p} \Phi_0^{\kappa, \lambda}(0, 0) \xrightarrow{\mathbb{P}} 0.$$

By a standard probability argument (see for instance [55]), it suffices to show that  $\mathbf{E} \left[ \phi \left( \frac{1}{p} \Phi_n^{\kappa, \lambda}(0, 0) \right) \right] - \mathbf{E} \left[ \phi \left( \frac{1}{p} \Phi_0^{\kappa, \lambda}(0, 0) \right) \right] \rightarrow 0$  for any bounded test function  $\phi$  that has bounded derivatives up to

the third order. To bound this difference, we will approximate the problems at  $(0,0)$ , that is,  $\Phi_n^{\kappa,\lambda}(0,0), \Phi_0^{\kappa,\lambda}(0,0)$  with the corresponding problems for positive  $\epsilon, \delta$ . To control this approximation, we further need to control the approximation error of  $h(\cdot)$ , using  $h_\delta(\cdot)$ , and the derivatives of  $h_\delta(\cdot)$ . This is achieved in Lemma B.2. On working out this argument, we obtain that

$$\begin{aligned}
& \left| \mathbf{E} \left[ \phi \left( \frac{1}{p} \Phi_n^{\kappa,\lambda}(0,0) \right) \right] - \mathbf{E} \left[ \phi \left( \frac{1}{p} \Phi_0^{\kappa,\lambda}(0,0) \right) \right] \right| \\
& \leq \left| \mathbf{E} \left[ \phi \left( \frac{1}{p} \Phi_n^{\kappa,\lambda}(\epsilon, \delta) \right) \right] - \mathbf{E} \left[ \phi \left( \frac{1}{p} \Phi_0^{\kappa,\lambda}(\epsilon, \delta) \right) \right] \right| + C \cdot \|\phi'\|_{L^\infty}(\delta + \epsilon), \quad \text{Lemma B.2} \\
& \leq \underbrace{\sum_{k=0}^{n-1} \left| \mathbf{E} \left[ \phi \left( \frac{1}{p} \Phi_{k+1}^{\kappa,\lambda}(\epsilon, \delta) \right) \right] - \mathbf{E} \left[ \phi \left( \frac{1}{p} \Phi_k^{\kappa,\lambda}(\epsilon, \delta) \right) \right] \right|}_{(i)} + C \cdot \|\phi'\|_{L^\infty}(\delta + \epsilon). \tag{B.14}
\end{aligned}$$

Above  $C$  involves universal constants and the scaled norms  $\|\theta_n^*\|^2/p, \|\theta_0^*\|^2/p$ , where these denote optimizers of the objective functions in  $\Phi_n^{\kappa,\lambda}(0,0), \Phi_0^{\kappa,\lambda}(0,0)$ . Thus, it suffices to control  $(i)$ , which we achieve by a Lindeberg argument. Denoting  $\mathbf{E}_x$  to be the expectation with respect to  $x$ , keeping all other random variables fixed, we note that

$$(i) \leq \|\phi'\|_{L^\infty} \frac{1}{p} \mathbf{E} \left| \mathbf{E}_{\tilde{x}_k} [\Phi_k^{\kappa,\lambda}(\epsilon, \delta)] - \mathbf{E}_{\tilde{x}_{k-1}} [\Phi_{k-1}^{\kappa,\lambda}(\epsilon, \delta)] \right|. \tag{B.15}$$

Define  $\ell_\delta(x, y) = h_\delta^2(\kappa - y \cdot x)$ . Then, as in [55, Eqn. 35], we can define the following quadratic approximation to the leave-one-out problem  $\Phi_{\setminus k}^{\kappa,\lambda}(\epsilon, \delta)$

$$\Psi_k(x) := \Phi_{\setminus k}^{\kappa,\lambda}(\epsilon, \delta) + \min_{\theta} \left\{ \frac{1}{2} (\theta - \theta_{\setminus k}^*)^\top H_{\setminus k}(\epsilon) (\theta - \theta_{\setminus k}^*) + \ell_\delta \left( \frac{1}{\sqrt{p}} x^\top \theta, y_k \right) \right\}. \tag{B.16}$$

With this notation, bounding the RHS of (B.15) breaks down to two tasks—controlling the error of quadratic approximation

$$(ii) := \max \left\{ |\Phi_k^{\kappa,\lambda}(\epsilon, \delta) - \Psi_k(x_k)|, |\Phi_{k-1}^{\kappa,\lambda}(\epsilon, \delta) - \Psi_k(\tilde{x}_k)| \right\} \tag{B.17}$$

and the error term

$$(iii) := \left| \mathbf{E}_{x_k} [\Psi_k(x_k)] - \mathbf{E}_{\tilde{x}_k} [\Psi_k(\tilde{x}_k)] \right| \tag{B.18}$$

Define the Moreau envelope to be

$$\mathcal{M}_k(t, \gamma_k) := \min_{s \in \mathbb{R}} \left\{ \ell_\delta(s, y_k) + \frac{(t - s)^2}{2\gamma_k} \right\}, \tag{B.19}$$

where the regularization parameter is defined to be

$$\gamma_k = \frac{1}{p} \mathbf{E}_{\mathbf{x}} \left[ \mathbf{x}^\top (H_{\setminus k}(\epsilon))^{-1} \mathbf{x} \right] \leq \epsilon^{-1}. \tag{B.20}$$

Due to the matching second moment of  $x_k$  and  $\tilde{x}_k$ , the above  $\gamma_k$  stays the same when  $\mathbf{x}$  is either  $x_k$  or  $\tilde{x}_k$  in distribution. Then, (iii) can be upper bounded by

$$\underbrace{\left| \mathbf{E}_{x_k}[\mathcal{M}_k(\frac{1}{\sqrt{p}}x_k^\top \theta_{\setminus k}^\star, \gamma_k)] - \mathbf{E}_{\tilde{x}_k}[\mathcal{M}_k(\frac{1}{\sqrt{p}}\tilde{x}_k^\top \theta_{\setminus k}^\star, \gamma_k)] \right|}_{(iv)} + (v), \quad (\text{B.21})$$

with

$$\begin{aligned} (v) := & \left| \mathbf{E}_{x_k}[\mathcal{M}_k(\frac{1}{\sqrt{p}}x_k^\top \theta_{\setminus k}^\star, \gamma_k)] - \mathcal{M}_k(\frac{1}{\sqrt{p}}x_k^\top \theta_{\setminus k}^\star, \gamma(x_k)) \right| \\ & + \left| \mathbf{E}_{\tilde{x}_k}[\mathcal{M}_k(\frac{1}{\sqrt{p}}\tilde{x}_k^\top \theta_{\setminus k}^\star, \gamma_k)] - \mathcal{M}_k(\frac{1}{\sqrt{p}}\tilde{x}_k^\top \theta_{\setminus k}^\star, \gamma(\tilde{x}_k)) \right| \end{aligned} \quad (\text{B.22})$$

where  $\gamma(x) := \frac{1}{p}x^\top (H_{\setminus k}(\epsilon))^{-1}x$ .

Thus, it suffices to bound (ii), (iv) and (v). This requires controlling the approximation error of  $h(\cdot)$  using  $h_\delta(\cdot)$ , derivatives of  $h_\delta(\cdot)$  and the Moreau envelope. We achieve these in Lemma B.1-B.2, and using these, we claim the following bounds,

$$(ii) \leq \frac{\text{poly}(\epsilon^{-1}, \delta^{-1})}{\sqrt{p}} \text{polylog}(p), \quad (\text{B.23})$$

$$(iv) \leq \frac{\text{poly}(\epsilon^{-1})}{\sqrt{p}} \text{polylog}(p), \quad (\text{B.24})$$

$$(v) \leq \frac{\epsilon^{-2}}{\sqrt{p}} \text{polylog}(p), \quad (\text{B.25})$$

We will prove these invoking Lemma B.2-B.1 and techniques from [55, Lemma 1, 2, 24]. Before we present the proofs, note that, together with (B.14), this implies that with proper choice of  $\epsilon, \delta = p^{-c_1}$

$$\left| \mathbf{E} \left[ \phi \left( \frac{1}{p} \Phi_n^{\kappa, \lambda}(0, 0) \right) \right] - \mathbf{E} \left[ \phi \left( \frac{1}{p} \Phi_0^{\kappa, \lambda}(0, 0) \right) \right] \right| \lesssim p^{-c_2}, \quad (\text{B.26})$$

with some  $c_1, c_2 > 0$ . The above is true since it can be shown that the constant  $C$  in (B.14) is  $O(1)$ , by arguments similar to (B.33). The rest of the proof thus focuses on establishing (B.23)–(B.25).

#### Proof of Eqn. (B.25)

*Proof of Eqn. (B.25).* To bound the term (v), it suffices to control  $\left| \mathbf{E}_{x_k}[\mathcal{M}_k(t, \gamma(x_k)) - \mathcal{M}_k(t, \gamma_k)] \right|$  with  $t = \frac{1}{\sqrt{p}}x_k^\top \theta_{\setminus k}^\star$ . First, calculate the partial derivative of  $\mathcal{M}_k(t, \gamma)$  w.r.t.  $\gamma$ ,

$$\left| \frac{\partial}{\partial \gamma} \mathcal{M}_k(t, \gamma) \right| = \frac{1}{2} |\ell'_\delta(s^\star, y_k)|^2 \quad (\text{B.27})$$

where  $s^\star$  satisfies  $\ell'_\delta(s^\star, y_k) + \frac{s^\star - t}{\gamma} = 0$  (for fixed  $t, \gamma$ ). By (B.63), the following upper bound holds

$$\left| \frac{\partial}{\partial \gamma} \mathcal{M}_k(t, \gamma) \right| \leq 2(2\kappa + \delta + |t|)^2. \quad (\text{B.28})$$

With the above, we know

$$\left| \mathbb{E}_{x_k} [\mathcal{M}_k(t, \gamma(x_k)) - \mathcal{M}_k(t, \gamma_k)] \right| = \left| \mathbb{E}_{x_k} \left[ \frac{\partial}{\partial \gamma} \mathcal{M}_k(t, \tilde{\gamma})(\gamma(x_k) - \gamma_k) \right] \right| \quad (\text{B.29})$$

$$\leq \sqrt{\mathbb{E}_{x_k} \left| \frac{\partial}{\partial \gamma} \mathcal{M}_k(t, \tilde{\gamma}) \right|^2 \cdot \mathbb{E}_{x_k} (\gamma(x_k) - \gamma_k)^2} \quad (\text{B.30})$$

$$\lesssim \sqrt{\left[ \kappa^4 + \delta^4 + \mathbb{E}_{x_k} \left( \frac{1}{\sqrt{p}} x_k^\top \theta_{\setminus k}^\star \right)^4 \right] \cdot \mathbb{E}_{x_k} \left( \frac{1}{p} x_k^\top (H_{\setminus k}(\epsilon))^{-1} x_k - \gamma_k \right)^2} \quad (\text{B.31})$$

$$\lesssim \sqrt{\left[ \kappa^4 + \delta^4 + \left\| \frac{1}{\sqrt{p}} \theta_{\setminus k}^\star \right\|^4 \right] \cdot \frac{\epsilon^{-2}}{p}} \quad (\text{B.32})$$

and hence it suffices to bound  $\left\| \frac{1}{\sqrt{p}} \theta_{\setminus k}^\star \right\|$ . Note that, in the definition of  $\gamma_k$  above,  $x$  has the same distribution as  $x_k$  so that the term  $E_{x_k} \left( \frac{1}{p} x_k^\top (H_{\setminus k}(\epsilon))^{-1} x_k - \gamma_k \right)^2$  can effectively be treated as variance of a chi-square random variable with  $p$  degrees of freedom, scaled by  $p$ . We know

$$\frac{\epsilon}{2} \|\theta_{\setminus k}^\star\|^2 \leq \mathcal{L}_{\setminus k}^{\kappa, \lambda}(\theta_{\setminus k}^\star; \epsilon, \delta) + p\lambda \leq \mathcal{L}_{\setminus k}^{\kappa, \lambda}(0; \epsilon, \delta) + p\lambda \leq n(\kappa + \delta)^2 + p\lambda. \quad (\text{B.33})$$

Putting things together, we have

$$\left| \mathbb{E}_{x_k} [\mathcal{M}_k(t, \gamma(x_k)) - \mathcal{M}_k(t, \gamma_k)] \right| \lesssim \epsilon^{-2} \frac{1}{\sqrt{p}} \quad (\text{B.34})$$

□

#### Proof of Eqn. (B.24)

*Proof of Eqn. (B.24).* The proof follows directly from Lemma B.1 and [55, Lemma 2] since Lemma B.1 verifies the needed condition needed in [55, Lemma 2]. □

#### Proof of Eqn. (B.23)

*Proof of Eqn. (B.23).* Define the following three minimizers

$$\theta^\star(x_k) = \arg \min_{\theta} \left\{ \mathcal{L}_{\setminus k}^{\kappa, \lambda}(\theta; \epsilon, \delta) + h_\delta^2(\kappa - y_k \frac{1}{\sqrt{p}} x_k^\top \theta) \right\} \quad (\text{B.35})$$

$$\tilde{\theta}(x_k) = \arg \min_{\theta} \left\{ \frac{1}{2}(\theta - \theta_{\setminus k}^\star)^\top H_{\setminus k}(\epsilon)(\theta - \theta_{\setminus k}^\star) + h_\delta^2(\kappa - y_k \frac{1}{\sqrt{p}} x_k^\top \theta) \right\} \quad (\text{B.36})$$

$$\theta_{\setminus k}^\star = \arg \min_{\theta} \mathcal{L}_{\setminus k}^{\kappa, \lambda}(\theta; \epsilon, \delta) \quad (\text{B.37})$$

To upper bound (ii), we need to control the quadratic approximation

$$\begin{aligned} & \left| \Phi_k^{\kappa, \lambda}(\epsilon, \delta) - \Psi_k(x_k) \right| \\ & \leq \max_{\theta \in \{\theta^\star(x_k), \tilde{\theta}(x_k)\}} \left\{ \mathcal{L}_{\setminus k}^{\kappa, \lambda}(\theta; \epsilon, \delta) - \mathcal{L}_{\setminus k}^{\kappa, \lambda}(\theta_{\setminus k}^\star; \epsilon, \delta) - \frac{1}{2}(\theta - \theta_{\setminus k}^\star)^\top H_{\setminus k}(\epsilon)(\theta - \theta_{\setminus k}^\star) \right\} \end{aligned} \quad (\text{B.38})$$



By a Taylor expansion up to the third order and the mean value theorem, the above expression can be bounded by

$$\begin{aligned} & \frac{1}{6} n \cdot \max_{i \neq k} \underbrace{|2h_\delta'''(t)h_\delta(t) + 6h_\delta''(t)h_\delta'(t)|_{t=\kappa-y_i \frac{1}{\sqrt{p}} x_i^\top \theta_m}}_{(b)} \cdot \underbrace{\left| \frac{1}{\sqrt{p}} x_i^\top (\theta - \theta_{\setminus k}^\star) \right|^3}_{(a)} \\ & + \frac{1}{6} \sum_{j=1}^p \lambda \underbrace{|h_\delta'''(t) - h_\delta'''(-t)|_{t=\theta_m[j]}}_{(c)} \cdot |\theta[j] - \theta_{\setminus k}^\star[j]|^3, \end{aligned} \quad (\text{B.39})$$

where  $\theta_m = (1 - \xi)\theta_{\setminus k}^\star + \xi\theta$  is an intermediate point. Here when we take  $\max_{i \neq k}$ , we slightly abuse the notation: for  $i = 1, \dots, k-1$ , (a) and (b) is as stated with  $x_i$ ; for  $i = k+1, \dots, n$ , (a) and (b) should have  $\tilde{x}_i$  substituting  $x_i$ . In the rest of the proof, when we control (a) and (b), the proof follows the same way with either  $\tilde{x}_i$  or  $x_i$ .

**Case 1:**  $\theta = \tilde{\theta}(x_k)$ . To handle the case of  $\theta = \tilde{\theta}(x_k)$ , note that

$$\tilde{\theta}(x_k) - \theta_{\setminus k}^\star = 2 \left[ h_\delta'(t) h_\delta(t) \right]_{t=\kappa-y_k \frac{1}{\sqrt{p}} x_k^\top \tilde{\theta}(x_k)} [H_{\setminus k}(\epsilon)]^{-1} \frac{1}{\sqrt{p}} y_k x_k. \quad (\text{B.40})$$

With this fact in mind, we continue to control each term (a), (b), (c).

**Term (a):**

$$\left| \frac{1}{\sqrt{p}} x_i^\top (\tilde{\theta}(x_k) - \theta_{\setminus k}^\star) \right| \quad (\text{B.41})$$

$$= 2 |h_\delta'(t) h_\delta(t)|_{t=\kappa-y_k \frac{1}{\sqrt{p}} x_k^\top \tilde{\theta}(x_k)} \cdot \left| \frac{1}{p} x_i^\top [H_{\setminus k}(\epsilon)]^{-1} x_k \right| \quad \text{by (B.40)} \quad (\text{B.42})$$

$$\lesssim |h_\delta(t)|_{t=\kappa-y_k \frac{1}{\sqrt{p}} x_k^\top \theta_{\setminus k}^\star} \frac{\epsilon^{-1} \text{polylog}(p)}{p^{0.5}} \quad (\text{B.43})$$

$$\lesssim \frac{\epsilon^{-1.5} \text{polylog}(p)}{p^{0.5}} \quad (\text{B.44})$$

where the second to last step uses two facts (1)  $|h_\delta'(t)| \leq 1$  and  $h_\delta^2(\kappa - y_k \frac{1}{\sqrt{p}} x_k^\top \theta_{\setminus k}^\star) \geq h_\delta^2(\kappa - y_k \frac{1}{\sqrt{p}} x_k^\top \tilde{\theta}(x_k)) + \frac{1}{2} (\tilde{\theta}(x_k) - \theta_{\setminus k}^\star)^\top H_{\setminus k}(\epsilon) (\tilde{\theta}(x_k) - \theta_{\setminus k}^\star) \geq h_\delta^2(\kappa - y_k \frac{1}{\sqrt{p}} x_k^\top \tilde{\theta}(x_k))$ , (2) [55, Lemma 10]. Recalling the estimate on  $\frac{1}{p} \|\theta_{\setminus k}^\star\|^2 \lesssim \epsilon^{-1}$  as in (B.33), we obtain the last step.

**Term (b):**

$$\begin{aligned} & |2h_\delta'''(t)h_\delta(t) + 6h_\delta''(t)h_\delta'(t)|_{t=\kappa-y_i \frac{1}{\sqrt{p}} x_i^\top \theta_m} \\ & \lesssim \delta^{-2}(|t| + \delta) \quad \text{here } t = \kappa - y_i \frac{1}{\sqrt{p}} x_i^\top \theta_m, \text{ by Lemma B.2} \\ & \lesssim \delta^{-2} \left\{ \kappa + \max \left\{ \left| \frac{1}{\sqrt{p}} x_i^\top \theta_{\setminus k}^\star \right|, \left| \frac{1}{\sqrt{p}} x_i^\top \tilde{\theta}(x_k) \right| \right\} \right\} \\ & \lesssim \delta^{-2} \left\{ \left| \frac{1}{\sqrt{p}} x_i^\top \theta_{\setminus k}^\star \right| + \left| \frac{1}{\sqrt{p}} x_i^\top (\tilde{\theta}(x_k) - \theta_{\setminus k}^\star) \right| \right\} \\ & \lesssim \delta^{-2} \left\{ \left| \frac{1}{\sqrt{p}} x_i^\top \theta_{\setminus k}^\star \right| + \frac{\epsilon^{-1.5} \text{polylog}(p)}{p^{1/2}} \right\} \quad \text{by (B.44)} \\ & \lesssim \delta^{-2} \left\{ \left| \frac{1}{\sqrt{p}} x_i^\top \theta_{\setminus \{k,i\}}^\star \right| + \left| \frac{1}{\sqrt{p}} x_i^\top (\theta_{\setminus k}^\star - \theta_{\setminus \{k,i\}}^\star) \right| + \frac{\epsilon^{-1.5} \text{polylog}(p)}{p^{1/2}} \right\} \\ & \lesssim \delta^{-2} \left\{ \epsilon^{-0.5} \vee \epsilon^{-1} \text{polylog}(p) + \frac{\epsilon^{-1.5} \text{polylog}(p)}{p^{1/2}} \right\}. \end{aligned} \quad (\text{B.45})$$

The last two steps use the following fact:

$$\begin{aligned}
& \mathcal{L}_{\setminus\{k,i\}}^{\kappa,\lambda}(\theta_{\setminus\{k,i\}}^{\star}; \epsilon, \delta) + h_{\delta}^2(\kappa - y_i \frac{1}{\sqrt{p}} x_i^{\top} \theta_{\setminus\{k,i\}}^{\star}) \\
& \geq \mathcal{L}_{\setminus\{k,i\}}^{\kappa,\lambda}(\theta_{\setminus k}^{\star}; \epsilon, \delta) + h_{\delta}^2(\kappa - y_i \frac{1}{\sqrt{p}} x_i^{\top} \theta_{\setminus k}^{\star}) \\
& \geq \mathcal{L}_{\setminus\{k,i\}}^{\kappa,\lambda}(\theta_{\setminus\{k,i\}}^{\star}; \epsilon, \delta) + \frac{\epsilon}{2} \|\theta_{\setminus k}^{\star} - \theta_{\setminus\{k,i\}}^{\star}\|^2 + h_{\delta}^2(\kappa - y_i \frac{1}{\sqrt{p}} x_i^{\top} \theta_{\setminus k}^{\star}) \quad \text{by strong convexity,}
\end{aligned} \tag{B.46}$$

and hence we have

$$\begin{aligned}
\frac{\epsilon}{2} \|\theta_{\setminus k}^{\star} - \theta_{\setminus\{k,i\}}^{\star}\|^2 & \leq h_{\delta}^2(\kappa - y_i \frac{1}{\sqrt{p}} x_i^{\top} \theta_{\setminus\{k,i\}}^{\star}) \leq \left( \kappa + \delta + \frac{1}{\sqrt{p}} \|\theta_{\setminus\{k,i\}}^{\star}\| \cdot \text{polylog}(p) \right)^2 \\
& \|\theta_{\setminus k}^{\star} - \theta_{\setminus\{k,i\}}^{\star}\|^2 \leq \epsilon^{-2} \text{polylog}(p)
\end{aligned} \tag{B.47}$$

as  $\frac{1}{p} \|\theta_{\setminus\{k,i\}}^{\star}\|^2 \lesssim \epsilon^{-1}$ .

**Term (c):** First, observe that by Lemma B.2,  $h_{\delta}''' \lesssim \delta^{-2}$ , thus we only need to control

$$\begin{aligned}
& \delta^{-2} \sum_{j=1}^p |\theta[j] - \theta_{\setminus k}^{\star}[j]|^3 \\
& \lesssim \delta^{-2} |h_{\delta}'(t) h_{\delta}(t)|_{t=\kappa - y_k \frac{1}{\sqrt{p}} x_k^{\top} \tilde{\theta}(x_k)}^3 \sum_{j=1}^p |h_j \frac{1}{\sqrt{p}} x_k|^3 \quad \text{where } h_j \in \mathbb{R}^n \text{ is the } j\text{-th column of } [H_{\setminus k}(\epsilon)]^{-1} \\
& \lesssim \delta^{-2} \epsilon^{-1.5} p \left( \frac{\|h_j\|}{p^{0.5}} \right)^3 \text{polylog } p \quad \text{since } \|h_j\| \leq \epsilon^{-1} \\
& \lesssim \frac{\delta^{-2} \epsilon^{-4.5} \text{polylog}(p)}{p^{0.5}}.
\end{aligned} \tag{B.48}$$

Putting the upper bounds on (a), (b) and (c) together, we effectively have (B.39) with  $\theta = \tilde{\theta}(x_k)$  is upper bounded by

$$\begin{aligned}
(\text{B.39}) & \lesssim \frac{\delta^{-2} \epsilon^{-5.5}}{p^{0.5}} \text{polylog}(p) \left( 1 \vee \frac{\epsilon^{-1} \text{polylog}(p)}{p^{0.5}} \right) + \frac{\delta^{-2} \epsilon^{-4.5}}{p^{0.5}} \text{polylog}(p) \\
& = \frac{\text{poly}(\epsilon^{-1}, \delta^{-1})}{p^{0.5}} \text{polylog}(p).
\end{aligned} \tag{B.49}$$

**Case 2:**  $\theta = \theta^{\star}(x_k)$ . Compared to **Case 1**, here we need an additional fact that controls the deviation

$$\|\theta^{\star}(x_k) - \tilde{\theta}(x_k)\|. \tag{B.50}$$

Due to the strong convexity of  $\mathcal{L}_k^{\kappa,\lambda}(\theta; \epsilon, \delta)$ , we know

$$\begin{aligned}
& \|\theta^\star(x_k) - \tilde{\theta}(x_k)\| \leq \epsilon^{-1} \|\nabla \mathcal{L}_k^{\kappa,\lambda}(\theta^\star(x_k); \epsilon, \delta) - \nabla \mathcal{L}_k^{\kappa,\lambda}(\tilde{\theta}(x_k); \epsilon, \delta)\| \\
& = \epsilon^{-1} \|\nabla \mathcal{L}_k^{\kappa,\lambda}(\tilde{\theta}(x_k); \epsilon, \delta)\| \\
& = \epsilon^{-1} \|\nabla \mathcal{L}_k^{\kappa,\lambda}(\tilde{\theta}(x_k); \epsilon, \delta) - \nabla \mathcal{L}_{\setminus k}^{\kappa,\lambda}(\theta_{\setminus k}^\star; \epsilon, \delta)\| \\
& = \epsilon^{-1} \|\nabla \mathcal{L}_{\setminus k}^{\kappa,\lambda}(\tilde{\theta}(x_k); \epsilon, \delta) + \nabla h_\delta^2(\kappa - y_k \frac{1}{\sqrt{p}} x_k^\top \tilde{\theta}(x_k)) - \nabla \mathcal{L}_{\setminus k}^{\kappa,\lambda}(\theta_{\setminus k}^\star; \epsilon, \delta)\| \\
& = \epsilon^{-1} \|\nabla \mathcal{L}_{\setminus k}^{\kappa,\lambda}(\tilde{\theta}(x_k); \epsilon, \delta) - H_{\setminus k}(\epsilon)(\tilde{\theta}(x_k) - \theta_{\setminus k}^\star) - \nabla \mathcal{L}_{\setminus k}^{\kappa,\lambda}(\theta_{\setminus k}^\star; \epsilon, \delta)\| \\
& \lesssim \epsilon^{-1} \left\| \sum_{i=1}^{k-1} [2h_\delta'''(t)h_\delta(t) + 6h_\delta''(t)h_\delta'(t)]_{t=\kappa-y_i \frac{1}{\sqrt{p}} x_i^\top \theta_m} \left( \frac{1}{\sqrt{p}} x_i^\top (\tilde{\theta}(x_k) - \theta_{\setminus k}^\star) \right)^2 y_i \frac{1}{\sqrt{p}} x_i \right. \\
& \quad \left. + \sum_{i=k+1}^n [2h_\delta'''(t)h_\delta(t) + 6h_\delta''(t)h_\delta'(t)]_{t=\kappa-y_i \frac{1}{\sqrt{p}} \tilde{x}_i^\top \theta_m} \left( \frac{1}{\sqrt{p}} \tilde{x}_i^\top (\tilde{\theta}(x_k) - \theta_{\setminus k}^\star) \right)^2 y_i \frac{1}{\sqrt{p}} \tilde{x}_i \right\| \\
& \quad + \sqrt{\sum_{j=1}^p \lambda^2 |h_\delta'''(t) - h_\delta'''(-t)|_{t=\theta_m[j]}^2 (\tilde{\theta}(x_k)[j] - \theta_{\setminus k}^\star[j])^4}
\end{aligned} \tag{B.51}$$

where  $\theta_m$  lies between  $\tilde{\theta}(x_k)$  and  $\theta_{\setminus k}^\star$ . Using the same arguments as in **Case 1** for terms (a) and (b), we know

$$\begin{aligned}
\alpha_i &:= \left| [2h_\delta'''(t)h_\delta(t) + 6h_\delta''(t)h_\delta'(t)]_{t=\kappa-y_i \frac{1}{\sqrt{p}} \tilde{x}_i^\top \theta_m} \left( \frac{1}{\sqrt{p}} \tilde{x}_i^\top (\tilde{\theta}(x_k) - \theta_{\setminus k}^\star) \right)^2 y_i \right| \\
&\lesssim \delta^{-2} \epsilon^{-0.5} \cdot \left( \frac{\epsilon^{-1.5} \text{polylog}(p)}{p^{0.5}} \right)^2.
\end{aligned} \tag{B.52}$$

Therefore (B.51) can be bounded as follows

$$\begin{aligned}
(\text{B.51}) &\leq \epsilon^{-1} \left\| \left[ \frac{1}{\sqrt{p}} x_1, \dots, \frac{1}{\sqrt{p}} x_{k-1}, \frac{1}{\sqrt{p}} \tilde{x}_{k+1}, \dots, \frac{1}{\sqrt{p}} \tilde{x}_n \right] \right\|_{\text{op}} \|\alpha\| \\
&\lesssim \epsilon^{-1} \sqrt{n} \cdot \delta^{-2} \epsilon^{-0.5} \left( \frac{\epsilon^{-1.5} \text{polylog}(p)}{p^{0.5}} \right)^2 \lesssim \frac{\delta^{-2} \epsilon^{-4.5}}{p^{0.5}} \text{polylog}(p)
\end{aligned} \tag{B.53}$$

where we use the fact  $\left\| \left[ \frac{1}{\sqrt{p}} x_1, \dots, \frac{1}{\sqrt{p}} x_{k-1}, \frac{1}{\sqrt{p}} \tilde{x}_{k+1}, \dots, \frac{1}{\sqrt{p}} \tilde{x}_n \right] \right\|_{\text{op}} = O_P(1)$ . For (B.52), we know from the argument in bounding term (c) in **Case 1** that

$$(\text{B.52}) \lesssim \delta^{-2} \sqrt{\epsilon^{-3} p \frac{1}{p^2} \text{polylog}(p)} \lesssim \frac{\delta^{-2} \epsilon^{-4.5}}{p^{0.5}} \text{polylog}(p). \tag{B.54}$$

Therefore, we have established that

$$\|\theta^\star(x_k) - \tilde{\theta}(x_k)\| \lesssim \frac{\delta^{-2} \epsilon^{-4.5}}{p^{0.5}} \text{polylog}(p). \tag{B.55}$$

□

Now we revisit the terms (a), (b), (c) in the case where  $\theta = \theta^\star(x_k)$ .

**Term (a):**

$$\begin{aligned}
& \left| \frac{1}{\sqrt{p}} x_i^\top (\theta^\star(x_k) - \theta_{\setminus k}^\star) \right| \\
& \leq \left| \frac{1}{\sqrt{p}} x_i^\top (\tilde{\theta}(x_k) - \theta_{\setminus k}^\star) \right| + \left| \frac{1}{\sqrt{p}} x_i^\top (\tilde{\theta}(x_k) - \theta^\star(x_k)) \right| \\
& \lesssim \frac{\epsilon^{-1.5}}{p^{0.5}} \text{polylog}(p) + \frac{\delta^{-2} \epsilon^{-4.5}}{p^{0.5}} \text{polylog}(p) \quad \text{by (B.44) and (B.55)}.
\end{aligned} \tag{B.56}$$

**Term (b):**

$$\begin{aligned}
& |2h_\delta'''(t)h_\delta(t) + 6h_\delta''(t)h_\delta'(t)|_{t=\kappa - y_i \frac{1}{\sqrt{p}} x_i^\top \theta_m} \\
& \lesssim \delta^{-2} \left\{ \kappa + \max \left\{ \left| \frac{1}{\sqrt{p}} x_i^\top \theta_{\setminus k}^\star \right|, \left| \frac{1}{\sqrt{p}} x_i^\top \theta^\star(x_k) \right| \right\} \right\} \\
& \lesssim \delta^{-2} \left\{ \left| \frac{1}{\sqrt{p}} x_i^\top \theta_{\setminus k}^\star \right| + \left| \frac{1}{\sqrt{p}} x_i^\top (\theta^\star(x_k) - \theta_{\setminus k}^\star) \right| \right\} \\
& \lesssim \delta^{-2} \left\{ \epsilon^{-0.5} \vee \epsilon^{-1} \text{polylog}(p) + \frac{\delta^{-2} \epsilon^{-4.5}}{p^{0.5}} \text{polylog}(p) \right\}.
\end{aligned} \tag{B.57}$$

**Term (c):**

$$\begin{aligned}
& \delta^{-2} \sum_{j=1}^p |\theta^\star(x_k)[j] - \theta_{\setminus k}^\star[j]|^3 \\
& \lesssim \delta^{-2} \sum_{j=1}^p |\theta^\star(x_k)[j] - \tilde{\theta}(x_k)[j]|^3 + |\tilde{\theta}(x_k)[j] - \theta_{\setminus k}^\star[j]|^3 \\
& \lesssim \delta^{-2} \left\{ \sum_{j=1}^p (\theta^\star(x_k)[j] - \tilde{\theta}(x_k)[j])^2 \right\}^{3/2} + \delta^{-2} \sum_{j=1}^p |\tilde{\theta}(x_k)[j] - \theta_{\setminus k}^\star[j]|^3 \\
& \lesssim \delta^{-2} \|\theta^\star(x_k) - \tilde{\theta}(x_k)\|^3 + \frac{\delta^{-2} \epsilon^{-4.5}}{p^{0.5}} \text{polylog}(p) \\
& \lesssim \frac{\delta^{-8} \epsilon^{-13.5}}{p^{1.5}} \text{polylog}(p) + \frac{\delta^{-2} \epsilon^{-4.5}}{p^{0.5}} \text{polylog}(p).
\end{aligned} \tag{B.58}$$

Again, putting the upper bounds on (a), (b) and (c) together, we have shown (B.39) with  $\theta = \theta^\star(x_k)$  is upper bounded by

$$\text{(B.39)} \lesssim \frac{\text{poly}(\epsilon^{-1}, \delta^{-1})}{p^{0.5}} \text{polylog}(p). \tag{B.59}$$

□

### B.3 Supporting Lemmas

Throughout the proof of Theorem 3.4, we rely on the following two lemmas, and the proof is complete on proving these.

**Lemma B.1** (Moreau envelope). *Assume that  $\delta < \frac{\kappa}{2}\epsilon$ , the following estimates on the Moreau envelope hold,*

$$\begin{aligned}
\mathcal{M}_k(t, \gamma_k) & \leq (\kappa + \delta + |t|)^2, \\
\mathcal{M}'_k(t, \gamma_k) & \leq 2(2\kappa + \delta + |t|).
\end{aligned}$$

*Proof of Lemma B.1.* For the zeroth order estimate, we have

$$0 \leq \mathcal{M}_k(t, \gamma_k) \leq \ell_\delta(t, y_k) = h_\delta^2(\kappa - y_k t) \leq (\kappa + \delta + |t|)^2. \quad (\text{B.60})$$

For the first order estimate, we have by the Envelope Theorem

$$\begin{aligned} \mathcal{M}'_k(t, \gamma_k) &= \ell'_\delta(s, y_k) |_{s=s^\star(t)} \\ &= -2h_\delta(\kappa - y_k s^\star) h'_\delta(\kappa - y_k s^\star) y_k \\ |\mathcal{M}'_k(t, \gamma_k)| &\leq 2(\kappa + \delta + |s^\star(t)|) \end{aligned} \quad (\text{B.61})$$

where  $s^\star(t)$  is the solution to the equation on  $s$ , for any fixed  $t$  (proximal map)

$$s + \gamma_k \ell'_\delta(s, y_k) = t. \quad (\text{B.62})$$

Due to the non-expansiveness of the proximal map, we have

$$|\mathcal{M}'_k(t, \gamma_k)| \leq 2(\kappa + \delta + |t| + |s^\star(0)|) \leq 2(\kappa + \delta + |t| + \kappa) \quad (\text{B.63})$$

where the last step uses the fact

$$\frac{s^\star(0)}{\gamma_k} = 2h_\delta(\kappa - y_k s^\star(0)) h'_\delta(\kappa - y_k s^\star(0)) y_k \quad (\text{B.64})$$

and if  $y_k s^\star(0) > \kappa$ , we will reach a contradiction  $2\delta < \kappa \epsilon < \frac{\kappa}{\gamma_k} \leq 2|h_\delta(\kappa - y_k s^\star(0)) h'_\delta(\kappa - y_k s^\star(0))| \leq 2\delta$ .  $\square$

**Lemma B.2** (Gaussian smoothing). *The following estimates hold true*

$$\begin{aligned} |h_\delta(t) - h(t)| &\leq \sqrt{\frac{2}{\pi}} \delta, \\ |h'_\delta(t)| &\leq 1, \\ |h''_\delta(t)| &\leq 2\delta^{-2}|t| + 3\sqrt{\frac{2}{\pi}} \delta^{-1}, \\ |h'''_\delta(t)| &\leq 6\delta^{-2}. \end{aligned}$$

*Proof of Lemma B.2.* For the zeroth order estimate, we have

$$|h_\delta(t) - h(t)| = \mathbf{E}[|h(t + \delta \mathbf{g}) - h(t)|] \leq \delta \mathbf{E}[|\mathbf{g}|] = \sqrt{\frac{2}{\pi}} \delta. \quad (\text{B.65})$$

For the first order estimate,

$$\begin{aligned} |h'_\delta(t)| &= \left| \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}\delta} e^{-\frac{(s-t)^2}{2\delta^2}} \cdot \frac{(t-s)}{\delta^2} \cdot h(s) ds \right| = \delta^{-1} |\mathbf{E}[\mathbf{g} h(t + \delta \mathbf{g})]| \\ &\leq \delta^{-1} |\mathbf{E}[\mathbf{g} h(t)]| + \delta^{-1} \mathbf{E}[|\mathbf{g}| \cdot |h(t + \delta \mathbf{g}) - h(t)|] \leq 1. \end{aligned} \quad (\text{B.66})$$

For the second order estimate,

$$|h''_\delta(t)| = \delta^{-2} |\mathbf{E}[(1 + \mathbf{g}^2)h(t + \delta \mathbf{g})]| \leq 2\delta^{-2}|t| + 3\sqrt{\frac{2}{\pi}} \delta^{-1}. \quad (\text{B.67})$$

For the third order estimate,

$$|h'''_\delta(t)| = \delta^{-3} |\mathbf{E}[(3\mathbf{g} + \mathbf{g}^3)h(t + \delta \mathbf{g})]| \leq 6\delta^{-2}. \quad (\text{B.68})$$

$\square$