A Precise High-Dimensional Asymptotic Theory for Boosting and Min-L_1-Norm Interpolated Classifiers

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

[1]University of Chicago
[2]Harvard University

January 31, 2020

Abstract

This paper establishes a precise high-dimensional asymptotic theory for Boosting on separable data, taking statistical and computational perspectives. We consider the setting where the number of features (weak learners) p scales with the sample size n, in an over-parametrized regime. On the statistical front, we provide an exact analysis of the generalization error of Boosting, when the algorithm interpolates the training data and maximizes an empirical L_1 margin. The angle between the Boosting solution and the ground truth is characterized explicitly. On the computational front, we provide a sharp analysis of the stopping time when Boosting approximately maximizes the empirical L_1 margin. Furthermore we discover that, the larger the margin, the smaller the proportion of active features (with zero initialization). At the heart of our theory lies a detailed study of the maximum L_1 margin, using tools from convex geometry. The maximum L_1 margin can be precisely described by a new system of non-linear equations, which we study using a novel uniform deviation argument. Preliminary numerical results are presented to demonstrate the accuracy of our theory.

1 Introduction

In classification tasks, modern machine learning models are complex enough to provide solutions that achieve zero training error, even for random labels. Prominent examples include AdaBoost/boosting, multi-layer neural networks, and kernel machines. However, among the many solutions with exactly zero training error, not all present good generalization properties. Empirically, it is commonly observed that practical optimization algorithms — even running on sufficiently over-parametrized models — usually favor certain "minimal" ways of "interpolating" the training data, which is conjectured to amount to good generalization. Different optimization algorithms favor distinct notions of "minimalism". In this paper, we investigate one particular notion of L_1 "minimalism" in the classification setting, induced by the celebrated AdaBoost/Boosting

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

algorithm, and provide a precise statistical and computational study in the over-parametrized regime when the data is *separable*.

Boosting and connections to max-L_1-margin. Boosting (Freund and Schapire, 1995, 1996), motivated from online learning, is arguably one of the most powerful machine learning tools. (Rosset et al., 2004; Zhang and Yu, 2005) established that for separable data, *Boosting Algorithms* with infinitesimal stepsize converge to the *min-L_1-norm* direction, when left to run until convergence. To be precise, imagine that we have n i.i.d. observations $\{x_i, y_i\}_{i=1}^n$ where $y_i \in \{\pm 1\}$ denotes the labels and $x_i \in \mathbb{R}^p$ forms the vector of features. If $\hat{\theta}_{\text{boost}}^{t,\eta}$ denotes the iterates from a *Boosting Algorithm* with stepsize η at time t, (Rosset et al., 2004; Zhang and Yu, 2005) establish that

$$\lim_{\eta \to 0} \lim_{t \to \infty} \hat{\theta}_{\text{boost}}^{t,\eta} / \|\hat{\theta}_{\text{boost}}^{t,\eta}\|_1 = \hat{\theta}_{n,\ell_1} \ , \tag{1.1}$$

where $\hat{\theta}_{n,\ell_1}$ denotes the following *min-L_1-norm interpolated classifier*

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

At the same time, it is not hard to see that the *min-L_1-norm interpolated classifier* agrees with the *max-L_1-margin direction* given by

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

whenever the max-L_1-margin κ_{n,ℓ_1} is positive. Thus, a thorough understanding of the high-dimensional behavior of *max-L_1-margin* promises to yield a deeper understanding of *Boosting Algorithms* when the number of features scales with n. Here, *Boosting Algorithms* refer to a general class of boosting algorithms, described in detail in Section 2.

This paper. Consider the following statistical setting: $x_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Lambda)$ where $\Lambda \in \mathbb{R}^{p \times p}$ is a diagonal covariance matrix. Assume that the conditional distribution $y_i | x_i \sim \text{Ber}(f(x_i^\top \theta_\star))$, namely

$$\mathbb{P}(y_i = +1 | x_i) = f(x_i^\top \theta_\star) \ , \tag{1.4}$$

where $f : \mathbb{R} \to [0,1]$ is a well-specified function, and $\theta_\star \in \mathbb{R}^p$ is unknown. Denote the joint distribution of $(x, y) \sim P$. Furthermore, consider the following asymptotic regime

$$p/n \to \psi \in (0, \infty). \tag{1.5}$$

This paper characterizes the following properties of *Boosting Algorithms* for separable data, in the aforementioned statistical setting:

Statistical. How large is the empirical L_1-margin of the Boosting solution? What is the angle between the Boosting solution $\hat{\theta}$ (min-L_1-norm interpolated classifier) and the truth $\theta_\star$? What are the generalization properties of Boosting?

Computational. How many iterations of the Boosting algorithm (precisely as a function of over-parametrization p/n) are required for an ϵ-approximation to the max-L_1-margin problem? What is the proportion of features selected by *Boosting Algorithms* when the training error vanishes?

We seek to understand these questions as the over-parametrization ratio ψ varies. To approach the problem, we leverage the connection laid out in (1.1), and study the max-L_1-margin κ_{n,ℓ_1}, in the asymptotic regime (1.5). This in turn leads to a precise understanding of *Boosting Algorithms* in such high-dimensional settings.

Related Work. Recently, the seminal paper (Montanari et al., 2019) studied the asymptotic properties of the max-L_2-margin in classification settings. Building on the Convex Gaussian Minimax Theorem (CGMT), their results characterize the almost sure limit for the margin, as well as its generalization error in the regime (1.5). Our technical analysis of the max-L_1-margin is largely inspired by (Montanari et al., 2019), however, differences in the induced geometry between the L_2 and L_1 constraints lead to significant additional challenges, as we shall see in Section 4.1. The *Convex Gaussian Min-Max Theorem* (CGMT) (Thrampoulidis et al., 2015, 2014) involves a tight version of a classical Gaussian comparison inequality by Gordon (Gordon, 1988) , and has been widely used in high-dimensional statistics and information theory, for instance, to study the asymptotic mean-squared error of regularized M-estimators (Thrampoulidis et al., 2018), to characterize the performance of the SLOPE estimator in sparse linear regression (Hu and Lu, 2019) and to establish performance guarantees for PhaseMax (Dhifallah et al., 2018).

Since its introduction in (Freund and Schapire, 1995, 1996), there has been a vast and expansive literature on Boosting. Due to space constraints, we are unable to provide a complete literature review. We mention, as much as possible, the most relevant literature on separable data. (Rosset et al., 2004; Zhang and Yu, 2005) established the connection between max-L_1-margin and the asymptotic limit of the Boosting trajectory. (Telgarsky, 2013) further studied the shrinkage technique (Zhang and Yu, 2005) on a variety of stepsize choices. We remark that our setting is different in nature from the high-dimensional Boosting literature where a notion of sparsity is assumed on the unknown parameter $\theta_\star$ (Buhlmann, 2006). It is crucial to note that, in our setting, the L_1 connection arises due to the nature of the AdaBoost/Boosting algorithm, rather than due to sparsity assumptions on $\theta_\star$. Here we investigate the min-L_1-norm interpolated classifier that characterizes the limit of the Boosting solution on separable data. In recent times, min-norm interpolated solutions and their statistical properties have been extensively studied — see (Belkin et al., 2018a,b; Liang and Rakhlin, 2019; Belkin et al., 2019; Hastie et al., 2019; Bartlett et al., 2019; Liang et al., 2019) for the regression problem, and (Montanari et al., 2019) for the classification problem.

The rest of the paper is organized as follows. Section 2 introduces some crucial ingredients that are heavily used through the rest of the paper. Section 3 presents our main results that is followed by a sketch of the proofs in Section 4, whereas the details are deferred to the Appendix.

2 Crucial Building Blocks

Our approach is inspired by (Montanari et al., 2019), where they derived the asymptotic properties of the Max-L_2-Margin. The L_1-margin exhibits a different asymptotic behavior compared to that of the L_2-margin, but shares similar technical building blocks.

Setup and assumptions. Throughout, we consider a sequence of problems $\{y(n), X(n), \theta_\star(n)\}_{n\geq 1}$, such that $y(n) \in \mathbb{R}^n, \theta_\star(n) \in \mathbb{R}^{p(n)}$ and $X(n) \in \mathbb{R}^{n\times p(n)}$, where the i-th row $x_i \sim \mathcal{N}(0, \Lambda(n))$ and the i-th entry of $y(n)$ satisfies $y_i|x_i \sim \text{Ber}(f(\langle \theta_{\star,n}, x_i\rangle))$; here, $\Lambda(n) \in \mathbb{R}^{p(n)\times p(n)}$ is a diagonal covariance matrix. We consider the asymptotic regime (1.5), that is, $p(n)/n \to \psi \in (0,\infty)$, and work under the following assumptions.

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*

$$\rho(n) := \left(\theta_\star(n)^\top \Lambda(n) \theta_\star(n)\right)^{1/2} \qquad \text{and} \qquad \bar{w}_i(n) := \sqrt{p}\frac{\sqrt{\lambda_i(n)}\theta_{\star,n}^\top e_{i,n}}{\rho(n)}, \tag{2.1}$$

where $e_{i,n}$ denotes the canonical vector in $\mathbb{R}^n$ with 1 in the i-th entry and 0 elsewhere. Assume that the empirical distribution of $\{(\lambda_i(n), \bar{w}_i(n))\}_{i=1}^{p(n)}$ converges to a probability distribution μ on $\mathbb{R}_{>0} \times \mathbb{R}$, in the Wasserstein-2 distance, that is,

$$\frac{1}{p}\sum_{i=1}^{p} \delta_{(\lambda_i, \bar{w}_i)} \overset{W_2}{\Rightarrow} \mu \ , \tag{2.2}$$

which equivalently means weak convergence and convergence of the second moments (see for instance, (Montanari et al., 2019; Villani, 2008)). In particular, this implies that $\int w^2 \mu(d\lambda, dw) = 1$ and that $\rho(n) \to \rho$, and we further assume that $0 < \rho < \infty$.

Assumption 3. *Finally, assume that*

$$\|\bar{w}(n)\|_\infty \le C', \qquad \text{and} \qquad \|\bar{w}(n)\|_1/p > C'' \tag{2.3}$$

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

In the sequel, we will suppress the dependence on n for simplicity of the exposition.

A useful function. We next introduce the following function $F_\kappa : \mathbb{R} \times \mathbb{R}_{\ge 0} \to \mathbb{R}_{\ge 0}$ defined for any $\kappa \ge 0$,

$$F_\kappa(c_1, c_2) := (\mathbb{E}[(\kappa - c_1 Y Z_1 - c_2 Z_2)])^{\frac{1}{2}} \quad \text{where} \begin{cases} Z_2 \perp (Y, Z_1) \\ Z_i \sim \mathcal{N}(0,1), \ i = 1, 2 \\ \mathbb{P}(Y = +1|Z_1) = 1 - \mathbb{P}(Y = -1|Z_1) = f(\rho \cdot Z_1) \end{cases}. \tag{2.4}$$

We will shortly see that the threshold for separability of the data can be explicitly described in terms of (2.4). Thus, $F_\kappa(\cdot,\cdot)$ can be viewed as a generalization of the phase transition boundary for the MLE derived in (Candès and Sur, 2018), for the special case of the logistic link.

A non-linear system of equations. The asymptotic theory of the max-L_1-margin crucially depends on the behavior of a new non-linear system of equations.

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

$$\begin{aligned} c_1 &= -\mathop{\mathbf{E}}_{(\Lambda,W,G)\sim\mathcal{Q}}\left(\frac{\Lambda^{1/2}W \cdot \mathbf{prox}_s\left(\Lambda^{1/2}G + \psi^{-1/2}[\partial_1 F_\kappa(c_1,c_2) - c_1 c_2^{-1}\partial_2 F_\kappa(c_1,c_2)]\Lambda^{1/2}W\right)}{\psi^{-1/2}c_2^{-1}\partial_2 F_\kappa(c_1,c_2)}\right) \\ c_1^2 + c_2^2 &= \mathop{\mathbf{E}}_{(\Lambda,W,G)\sim\mathcal{Q}}\left(\frac{\mathbf{prox}_s\left(\Lambda^{1/2}G + \psi^{-1/2}[\partial_1 F_\kappa(c_1,c_2) - c_1 c_2^{-1}\partial_2 F_\kappa(c_1,c_2)]\Lambda^{1/2}W\right)}{\psi^{-1/2}c_2^{-1}\partial_2 F_\kappa(c_1,c_2)}\right)^2 \\ 1 &= \mathop{\mathbf{E}}_{(\Lambda,W,G)\sim\mathcal{Q}}\left|\frac{\mathbf{prox}_s\left(\Lambda^{1/2}G + \psi^{-1/2}[\partial_1 F_\kappa(c_1,c_2) - c_1 c_2^{-1}\partial_2 F_\kappa(c_1,c_2)]\Lambda^{1/2}W\right)}{\psi^{-1/2}c_2^{-1}\partial_2 F_\kappa(c_1,c_2)}\right|. \end{aligned} \tag{2.5}$$

Here, the expectation is over $(\Lambda, W, G) \sim \mu \otimes \mathcal{N}(0,1) =: \mathcal{Q}$ *with* μ *defined as in* (2.2)*, and the* $\mathbf{prox}_s(\cdot)$ *function is given by*

$$\mathbf{prox}_\lambda(t) = \arg\min_s \left\{ \lambda|s| + \frac{1}{2}(s-t)^2 \right\} = \mathrm{sign}(t)\,(|t| - \lambda)_+ \quad , \tag{2.6}$$

the proximal mapping operator of the L_1 *norm (Parikh and Boyd, 2014).*

Note that Λ denotes both the random variable in (2.5) and the covariance matrix in Assumption 1. Such overload of notation will prove useful in the technical derivations.

We emphasize that the equation system (2.5) differs significantly from that considered in the case of the L_2 geometry (Montanari et al., 2019; Shcherbina and Tirozzi, 2003; Gardner, 1988). Analogous systems arise in the study of high-dimensional statistical models in the proportional regime (1.5); here, the most relevant ones are the analysis of the MLE (Sur and Candès, 2019) and convex regularized estimators (Salehi et al., 2019) for logistic regression.

Uniqueness. It is insightful to understand when this system (2.5) admits a unique solution—observe that this is governed by the over-parametrization ratio ψ and the distribution $\mathcal{Q}$. To make it precise, introduce the constants

$$\zeta = \left(\mathop{\mathbf{E}}_{(\Lambda,W)\sim\mu} |\Lambda^{-1/2} W| \right)^{-1} \qquad \text{and} \qquad \omega = \left(\mathop{\mathbf{E}}_{(\Lambda,W)\sim\mu} \left[(1 - \zeta^2 \Lambda^{-1})^2 W^2 \right] \right)^2 . \tag{2.7}$$

In a similar spirit as in (Montanari et al., 2019), define the functions $\psi_+(\kappa) : \mathbb{R}_{>0} \to \mathbb{R}$, $\psi_- : \mathbb{R}_{>0} \to \mathbb{R}$ and $\psi^\downarrow(\kappa) : \mathbb{R}_{>0} \to \mathbb{R}_{\geq 0}$ as follows

$$\psi_+(\kappa) = \begin{cases} 0 & \text{if } \partial_1 F_\kappa(\zeta, 0) > 0 \\ \partial_2^2 F_\kappa(-\zeta, 0) - \omega^2 \partial_1^2 F_\kappa(-\zeta, 0) & \text{if o.w.} \end{cases} , \tag{2.8}$$

$$\psi_-(\kappa) = \begin{cases} 0 & \text{if } \partial_1 F_\kappa(-\zeta, 0) > 0 \\ \partial_2^2 F_\kappa(\zeta, 0) - \omega^2 \partial_1^2 F_\kappa(\zeta, 0) & \text{if o.w.} \end{cases} , \tag{2.9}$$

$$\psi^\downarrow(\kappa) = \max\{\psi^\star(0), \psi_+(\kappa), \psi_-(\kappa)\}, \tag{2.10}$$

where $\psi^\star(0)$ is given by $\psi^\star(0) = \min_{c\in\mathbb{R}} F_0^2(c, 1)$.

Proposition 2.1. *For any* $\psi > \psi^\downarrow(\kappa)$*, under Assumptions 1-3, the system of equations* (2.5) *admits a unique solution* $(c_1, c_2, s) \in \mathbb{R} \times \mathbb{R}_{>0} \times \mathbb{R}_{>0}$.

A second useful function. Finally, for any $\psi > \psi^\downarrow(\kappa)$, define $T : (\psi, \kappa) \to \mathbb{R}$ as

$$T(\psi, \kappa) := \psi^{-1/2} \left[F_\kappa(c_1, c_2) - c_1 \partial_1 F_\kappa(c_1, c_2) - c_2 \partial_2 F_\kappa(c_1, c_2) \right] - s, \tag{2.11}$$

where $c_1(\psi, \kappa), c_2(\psi, \kappa), s(\psi, \kappa)$ forms the unique solution to (2.5).

Proposition 2.2. *Under the assumptions of Proposition 2.1, the function* $(\psi, \kappa) \to T(\psi, \kappa)$ *is continuous in the domain* $\{(\psi, \kappa) : \psi > \psi^\downarrow(\kappa)\}$ *and strictly increasing with respect to* κ*, and strictly decreasing with respect to* ψ.

Separability. (Montanari et al., 2019, Theorem 1) established that the data is linearly separable when the limiting ratio $\psi > \psi^\star(0)$. We restrict ourselves to this separable regime.

Boosting algorithm. For convenience of the readers, we describe here the general *Boosting Algorithms* we work with in the rest of the exposition. We begin by briefing the steps involved in AdaBoost (Freund and Schapire, 1996, 1995). Suppose that each weak learner outputs a binary decision $X_{ij} = x_i[j] \in \{-1, +1\}$ and $y_i \in \{-1, +1\}$. AdaBoost consider the following updates:

- Initialize: $\eta_0 = 1/n \cdot \mathbf{1}_n \in \Delta_n$, $\theta_0 = 0$. Here, Δ_n refers to the standard probability simplex given by $\Delta_n := \{\boldsymbol{p} \in [0,1]^n : \sum_{i=1}^n p_i = 1\}$.
- At time $t \geq 0$:
  - Feature Selection: $v_{t+1} := \arg\min_{v \in \{\pm e_j\}_{j\in[p]}} \sum_{i\in[n]} \eta_t[i] \cdot I_{y_i x_i^\top v \leq 0}$;
  - Adaptive Stepsize α_t: $\alpha_t = \frac{1}{2}\log\left(\frac{1-\sum_{i\in[n]}\eta_t[i]\cdot I_{y_i x_i^\top v_{t+1}\leq 0}}{\sum_{i\in[n]}\eta_t[i]\cdot I_{y_i x_i^\top v_{t+1}\leq 0}}\right)$;
  - Coordinate Update: $\theta_{t+1} = \theta_t + \alpha_t \cdot v_{t+1}$;
  - Weight Update: $\eta_{t+1}[i] \propto \eta_t[i]\exp(-\alpha_t y_i x_i^\top v_{t+1})$, normalized such that $\eta_{t+1} \in \Delta_n$.
- Terminate after T steps, and output the vector θ_T.

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

$$v_{t+1} := \arg\max_{v \in \{e_j\}_{j\in[p]}} |\eta_t^\top Z e_j|, \quad \alpha_t := \eta_t^\top Z v_{t+1}, \tag{2.12}$$

where $Z = y \circ X \in \mathbb{R}^{n\times p}$.

3 Main Results

Our results may be broadly arranged along two veins: understanding the behavior of the max-L_1-margin, and that of *Boosting Algorithms*.

Asymptotics for Max-L_1-margin. Recall the definition of the max-L_1-margin from (1.3). For any $\psi > \psi^\star(0)$, define the constant

$$\kappa_\star(\psi, \mu) = \inf\{\kappa \geq 0 \ : \ T(\psi, \kappa) = 0\} \ . \tag{3.1}$$

Note that due to the definition of $\psi^\downarrow(\kappa)$, the region $\{(\psi,\kappa) : \psi > \psi^\downarrow(\kappa)\}$ contains $\{(\psi,\kappa) : T(\psi,\kappa) = 0\}$, which guarantees that (3.1) is well-defined.

Theorem 3.1. *Suppose Assumptions 1-3 hold and that $\psi > \psi^\star(0)$. Then the max-L_1-margin converges almost surely to $\kappa_\star(\psi,\mu)$, when appropriately rescaled by $\sqrt{p}$, that is,*

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

Informally, the max-L_1-margin is of the order $1/\sqrt{p}$, and Theorem 3.1 pins down the exact asymptotic value. Such a precise characterization directly yields insights into both statistical and computational properties of *Boosting Algorithms* in high dimensions, as we shall see shortly.

Although Theorem 3.1 provides a sharp description of the max-L_1-margin asymptotics, it is of natural interest to understand how the associated min-L_1-interpolated classifier (1.2) performs in terms of generalization. To this end, given a pair (ψ, μ) such that $\psi > \psi^{\downarrow}(\kappa_\star(\psi,\mu))$, define

$$\mathrm{Err}_\star(\psi,\mu) = \mathbb{P}\left(c_1^\star Y Z_1 + c_2^\star Z_1 < 0\right) \ , \tag{3.3}$$

where $c_i^\star := c_i(\psi, \kappa_\star(\psi,\mu))$, $i = 1, 2$, forms the unique solution to (2.5), $\kappa_\star(\cdot,\cdot)$ is given by (3.1) and (Y, Z_1, Z_2) follows the joint distribution specified in (2.4).

Theorem 3.2. *Under the assumptions of Theorem 3.1, the generalization error of the min-L_1-interpolated classifier $\hat{\theta}_{n,\ell_1}$, defined in (1.2), converges almost surely to* $\mathrm{Err}_\star(\psi,\mu)$, *that is, for a new data point* $(\mathbf{x}, \mathbf{y}) \sim P$,

$$\lim_{n\to\infty} \mathbb{P}_{(\mathbf{x},\mathbf{y})\sim P}\left(\mathbf{y}\cdot \mathbf{x}^\top \hat{\theta}_{n,\ell_1} < 0\right) \overset{\text{a.s.}}{=} \mathrm{Err}_\star(\psi,\mu) \ . \tag{3.4}$$

In view of the connection between *Boosting Algorithms* and the min-L_1-norm interpolated solution on separable data, Theorem 3.2 provides an exact quantification of the generalization behavior of *Boosting Algorithm* solutions with appropriately chosen learning rates. This result complements the classical empirical margin upper bounds on the generalization error (Koltchinskii and Panchenko, 2002). In addition, towards establishing these results, one can also show that the angle between the interpolated solution $\hat{\theta}_{n,\ell_1}$ and the target $\theta_\star$ converges to $\arccos(c_1^\star/\sqrt{(c_1^\star)^2 + (c_2^\star)^2})$.

Boosting in high dimensions. Consider the *Boosting Algorithm* described in Section 2, Eqn. 2.12. We have the following characterization.

Theorem 3.3. *Consider the separable case with a positive margin $\kappa_\star(\psi,\mu) > 0$. Under the assumptions of Theorem 3.1, with a suitably chosen learning rate (specified in Cor. 4.1), the sequence of iterates $\{\hat{\theta}^t\}_{t\ge 1}$ obtained from the* Boosting Algorithm *obeys the following property: for any $0 < \epsilon < 1$, when the number of iterations t satisfies*

$$t \ge T_\epsilon(p) \quad \textit{with} \lim_{n\to\infty} \frac{T_\epsilon(p)}{p\log^2 p} \overset{\text{a.s.}}{=} \frac{12\epsilon^{-2}}{\kappa_\star^2(\psi,\mu)} \ , \tag{3.5}$$

the solution $\hat{\theta}^t/\|\hat{\theta}^t\|_1$ forms an $(1-\epsilon)$-approximation to the Min-L_1-Interpolated Classifier a.s.,

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

For separable data that has a large and comparable number of samples and features, Theorem 3.3 directly informs a stopping rule for *Boosting Algorithms* so as to ensure (approximately) the generalization error given by (3.4). In addition, for any numerical accuracy ϵ, the stopping time $T_\epsilon(p)$ has a sharp asymptotic characterization (even in terms of constants).

Proportion of Activated Features for AdaBoost. AdaBoost/Boosting chooses features (weak learners) adaptively. To better understand the classifiers produced by such algorithms, we study the proportion of features that are activated when the training error vanishes.

Theorem 3.4. *Let $S_0(p)$ denote the number of features selected the first time t when the* Boosting Algorithm *achieves zero training error (with an initialization of $\hat{\theta}^0 = 0$), in the sense that,*

$$S_0(p) := \#\left\{j \in [p] : \hat{\theta}_j^t \neq 0\right\} \ , \qquad \textit{where} \qquad \frac{1}{n}\sum_{i=1}^{n} I_{y_i x_i^\top \hat{\theta}^t \leq 0} = 0. \tag{3.6}$$

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

$$\limsup_{p\to\infty} \frac{S_0(p)}{p \log^2 p} \leq \frac{12}{\kappa_\star^2(\psi,\mu)} \wedge 1, \quad a.s. \tag{3.7}$$

In other words, the larger the margin $\kappa_\star(\psi,\mu)$, the sparser the solution (with zero training error), with at most a $\frac{12}{\kappa_\star^2(\psi,\mu)} \wedge 1$ proportion of active features in the limit.

Numeric Validation of Theory. We proceed to test the validity of our theory on simulated datasets. Consider a grid of values for the over-parametrization ratio $\psi \in \Psi$. Here the covariance $\Lambda(n) = I$ (identity matrix), $\rho = 1$ and y_i's are generated from the logistic model. For each $\psi \in \Psi$, we generate multiple samples of size $n = 400$, and approximate the max-L_1-margin by (a) the solution to the corresponding linear program (LP); the blue points in Figure 1(a) depict these values when scaled by $\sqrt{p}$, and, (b) the asymptotic value predicted by the analytic formula (3.2); the red points in Figure 1(a) represent these values. Calculating our theoretical predictions involves solving (2.5), for which we approximate integrals via Monte-Carlo sums (5000 samples). Figure 1(b) compares the corresponding out-of-sample prediction error: the blue points show the generalization error $\mathbb{P}_{\mathbf{x},\mathbf{y}}\left(\mathbf{y}\cdot\mathbf{x}^\top\hat{\theta}_n < 0\right)$, when $\hat{\theta}_n$ is calculated from the LP, whereas the red points depict the asymptotic value $\mathrm{Err}_\star(\psi,\mu)$ predicted by our theory. Observe that in both cases, the points align remarkably well, particularly above a threshold for ψ, when the data is separable.

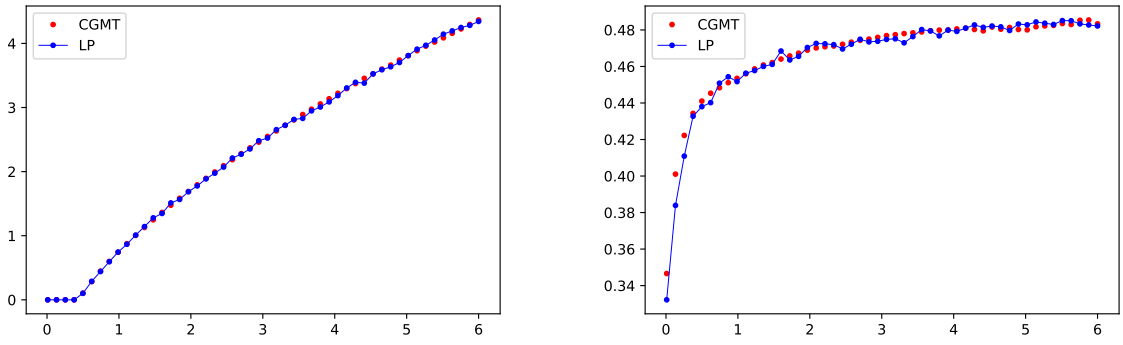

Figure 1: x-axis: Ratio p/n. y-axis: (a) Left: Max-L_1-margin, (b) Right: Generalization error

Now, note that if κ_{n,ℓ_2} denotes the usual L_2 margin, it is easy to see that

$$\kappa_{n,\ell_2} \leq \sqrt{p} \cdot \kappa_{n,\ell_1} \ . \tag{3.8}$$

The curious reader may wonder whether the L_1-margin, when appropriately scaled, differs significantly from the L_2-margin investigated in (Montanari et al., 2019). To study this, we consider the same setting as in (Montanari et al., 2019, Fig. 1, case of $\beta = 1$), and plot the max-L_1-margin limit given by (3.2). Evidently, the range for the ℓ_1-margin is roughly twice the size of that for the ℓ_2 case (Montanari et al., 2019), suggesting that these behave differently, even after appropriate scaling.

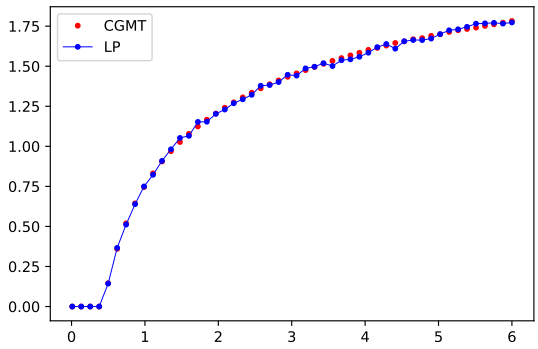

Figure 2: x-axis: varying ratio $\psi := p/n$. y-axis: $\kappa_\star/\sqrt{\psi}$.

4 Main Derivations

The proofs of Theorems 3.1 and 3.2 rely on the *Convex Gaussian Min-Max Theorem* (CGMT) (Thrampoulidis et al., 2015, 2014), a refinement of Gordon's classical Gaussian comparison inequality (Gordon, 1988). Our analysis builds heavily upon the seminal work of (Montanari et al., 2019) that characterized the max-L_2-margin using CGMT-based techniques. However, this approach cannot be adapted directly to the L_1 case, and requires establishing a novel and possibly stronger form of uniform deviation result (detailed in Step 3 below), which might be of standalone interest. Here, we provide a sketch of the main proof ideas, highlighting along the way the differences and subtleties between the L_1 and L_2 formulations.

4.1 Proofs of Theorems 3.1 and 3.2

Step 1: $\sqrt{p}$-rescaling of L_1 ball. To begin with, define

$$\begin{aligned}\xi_{\psi,\kappa}^{(n,p)} &:= \min_{\|\theta\|_1 \le \sqrt{p}} \max_{\|\lambda\|_2 \le 1, \lambda \ge 0} \frac{1}{\sqrt{p}} \lambda^T (\kappa \mathbf{1} - (y \odot X)\theta) \\ &= \min_{\|\theta\|_1 \le \sqrt{p}} \frac{1}{\sqrt{p}} \|(\kappa \mathbf{1} - (y \odot X)\theta)_+\|_2 \ .\end{aligned} \tag{4.1}$$

It is not hard to see that

$$\begin{aligned}\xi_{\psi,\kappa}^{(n,p)} = 0, \ \text{ if and only if } \ \kappa \le p^{1/2} \cdot \kappa_{\ell_1}\left(\{x_i, y_i\}_{i=1}^n\right) \ , \\ \xi_{\psi,\kappa}^{(n,p)} > 0, \ \text{ if and only if } \ \kappa > p^{1/2} \cdot \kappa_{\ell_1}\left(\{x_i, y_i\}_{i=1}^n\right) \ .\end{aligned} \tag{4.2}$$

Thus, to study the rescaled max-L_1-margin, it suffices to examine the value of $\xi^{(n,p)}_{\psi,\kappa}$.

Recall the diagonal covariance Λ and define $z_i := \Lambda^{-1/2} x_i \ \forall i \in [n]$. Since $\rho_p = \|\Lambda^{1/2}\theta_\star\|$,

$$x_i^\top \theta_\star = z_i^\top \Lambda^{1/2}\theta_\star = \rho_p \cdot z_i^\top w, \ \text{ where } w := \Lambda^{1/2}\theta_\star / \|\Lambda^{1/2}\theta_\star\| \ . \tag{4.3}$$

Hence, we can express $y \odot X$ as

$$y \odot X = (y \odot Z)\Lambda^{1/2} = (y \odot Z(\Pi_w + \Pi_{w^\perp}))\,\Lambda^{1/2} \overset{d}{=} \left((y \odot z)w^\top + Z\Pi_{w^\perp}\right)\Lambda^{1/2},$$

where $z \in \mathbb{R}^n$ and $Z \in \mathbb{R}^{n\times p}$ has independent standard Gaussian entries.

Eqn. (4.1) then reduces to

$$\xi^{(n,p)}_{\psi,\kappa}(z,Z) := \min_{\|\theta\|_1 \le \sqrt{p}} \max_{\|\lambda\|_2 \le 1, \lambda \ge 0} \frac{1}{\sqrt{p}} \lambda^T \left(\kappa \mathbf{1} - (y \odot z)\langle w, \Lambda^{1/2}\theta\rangle\right) - \frac{1}{\sqrt{p}} \lambda^T Z \Pi_{w^\perp}(\Lambda^{1/2}\theta) \ . \tag{4.4}$$

This step follows from (Montanari et al., 2019, Section 4.3), with a $\sqrt{p}$-rescaled L_1 ball constraint. In our setting, the rescaling is necessary so as to ensure a well-defined limit for the max-L_1-margin.

Step 2: reduction to Gordon's problem. Due to the min-max form of (4.4), one can use Gordon's Gaussian comparison inequality (Thrampoulidis et al., 2015, 2014; Gordon, 1988) to further simplify the problem. To this end, introduce the following "de-coupled" optimization problem

$$\begin{aligned}
&\hat{\xi}^{(n,p)}_{\psi,\kappa}(z,\tilde{z},g) \\
&:= \min_{\|\theta\|_1 \le \sqrt{p}} \max_{\|\lambda\|_2 \le 1, \lambda \ge 0} \frac{1}{\sqrt{p}} \lambda^T \left(\kappa \mathbf{1} - (y \odot z)\langle w, \Lambda^{1/2}\theta\rangle - \tilde{z}\|\Pi_{w^\perp}(\Lambda^{1/2}\theta)\|_2\right) + \frac{1}{\sqrt{p}}\|\lambda\|_2 \langle g, \Pi_{w^\perp}(\Lambda^{1/2}\theta)\rangle \\
&= \left[\min_{\|\theta\|_1 \le \sqrt{p}} \frac{1}{\sqrt{p}} \left\| \left(\kappa \mathbf{1} - (y \odot z)\langle w, \Lambda^{1/2}\theta\rangle - \tilde{z}\|\Pi_{w^\perp}(\Lambda^{1/2}\theta)\|_2\right)_+ \right\|_2 + \frac{1}{\sqrt{p}} \left\langle \Pi_{w^\perp}(g), \Lambda^{1/2}\theta \right\rangle \right]_+ ,
\end{aligned} \tag{4.5}$$

where $z, \tilde{z} \in \mathbb{R}^n$ and $g \in \mathbb{R}^p$ are independent isotropic Gaussian vectors. By CGMT (Thrampoulidis et al., 2015, Theorem 3) (see Theorem A.1 in the Appendix), we have

$$\mathbb{P}\left(\xi^{(n,p)}_{\psi,\kappa}(z,Z) \le t | y, z\right) \le 2\mathbb{P}\left(\hat{\xi}^{(n,p)}_{\psi,\kappa}(z,\tilde{z},g) \le t | y, z\right) \tag{4.6}$$

$$\mathbb{P}\left(\xi^{(n,p)}_{\psi,\kappa}(z,Z) \ge t | y, z\right) \le 2\mathbb{P}\left(\hat{\xi}^{(n,p)}_{\psi,\kappa}(z,\tilde{z},g) \ge t | y, z\right). \tag{4.7}$$

Marginalizing over y and z, this suggests that it suffices to study (4.5).

Step 3: large n,p limit, new uniform deviation result. Recall the function $F_\kappa(\cdot,\cdot)$ from (2.4), and define the empirical version

$$\widehat{F}_\kappa(c_1,c_2) := \left(\widehat{\mathbf{E}}_n[(\kappa - c_1 Y Z_1 - c_2 Z_2)_+^2]\right)^{1/2} \ , \tag{4.8}$$

where $\hat{\mathbf{E}}_n(f(Y,Z_1,Z_2))$ means that $Y, Z_1, Z_2 \in \mathbb{R}^n$ with entries $(Y_i, Z_{1,i}, Z_{2,i})$ arising from the joint distribution specified in (2.4), whereas $\hat{\mathbf{E}}_n$ denotes the expectation with respect to the empirical

distribution of $\{(Y_i, Z_{1,i}, Z_{2,i})\}_{i=1}^n$. Then with $\lambda = \mathrm{diag}(\Lambda)$ denoting the vectorized Λ, we can express $\hat{\xi}_{\psi,\kappa}^{(n,p)}(z, \tilde{z}, g)$ as the positive part of the following

$$\hat{\xi}_{\psi,\kappa}^{(n,p)}(\lambda, w, g) := \min_{\|\theta\|_1 \le \sqrt{p}} \left[\psi^{-1/2} \widehat{F}_\kappa \left(\langle w, \Lambda^{1/2}\theta \rangle, \|\Pi_{w^\perp}(\Lambda^{1/2}\theta)\|_2 \right) + \frac{1}{\sqrt{p}} \left\langle \Pi_{w^\perp}(g), \Lambda^{1/2}\theta \right\rangle \right] . \tag{4.9}$$

We seek to study (4.9) in the large sample and feature limits $n, p \to \infty$ with $p/n \to \psi$. On taking limits naively, one can reach the following infinite-dimensional convex problem,

$$\tilde{\xi}_{\psi,\kappa}^{(\infty,\infty)}(\Lambda, W, G) := \min_{\|h\|_{L_1(\mathcal{Q})} \le 1} \left[\psi^{-1/2} F_\kappa \left(\langle w, \Lambda^{1/2} h \rangle_{L_2(\mathcal{Q})}, \|\Pi_{w^\perp}(\Lambda^{1/2} h)\|_{L_2(\mathcal{Q})} \right) + \left\langle \Pi_{w^\perp}(G), \Lambda^{1/2} h \right\rangle_{L_2(\mathcal{Q})} \right]. \tag{4.10}$$

Here, the optimization variable is the set of function $\{h : \mathbb{R}^3 \to \mathbb{R}, h \in \mathcal{L}^2(\mathcal{Q})\}$, where $\mathcal{Q} = \mu \otimes \mathcal{N}(0,1)$ with μ defined as in (2.2).

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,

$$\lim_{p \to \infty, p/n(p) \psi} \hat{\xi}_{\psi,\kappa}^{(n,p)}(\lambda, w, g) \overset{\text{a.s.}}{=} \tilde{\xi}_{\psi,\kappa}^{(\infty,\infty)}(\Lambda, W, G). \tag{4.11}$$

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

Our *technical innovation* lies in the development of (4.11), 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 L_2-constraint $\|\theta\|_2 \le 1$ (the setting of (Montanari et al., 2019)) and the L_1-constraint $\|\theta\|_1 \le \sqrt{p}$ (our setting), $|c_1| \le M$ for all $p(n), n$ and some constant $M > 0$; for the L_1 case, this follows by noting that

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

by Assumption 3. Turning to the second variable, we see that under our L_1-constraint, c_2 may potentially grow as $\sqrt{p}$ whereas it remains bounded when the L_2-norm of θ is bounded. Naturally, the unbounded region for c_2 creates significant additional challenges in establishing (4.11).

To address this issue, we proceed as follows: (a) In our first and key step, we discover a crucial self-normalizing property of the partial derivatives $\partial_i \hat{F}(\kappa)$, using which we establish that the empirical partial derivatives converge uniformly to the corresponding derivatives $\partial_i F(\kappa), i = 1, 2$, over an unbounded region for c_2 (see Lemma A.2). (a) We then establish that the "empirical fixed point (fp) equations" obtained by analyzing the KKT conditions for (4.9)[1] converge uniformly (over an unbounded region for c_2), to the corresponding "fp equations obtained from the KKT conditions for (4.10)".[2] The convergence here is in the sense of (A.12). The analysis uses the key Lemma A.2. See Step 4 for description of these KKT equations. (c) Leveraging (b), we show that any solution $(\hat{c}_1, \hat{c}_2, \hat{s})$ of the empirical fp equations converges to the unique solution $(c_1^\star, c_2^\star, s^\star)$ of the fp equations from (4.10). See Appendix A.3 for uniqueness of the solution. (d) Now, (4.9) can

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

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

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 (4.10) with $s^\star$ and $F_\kappa, \partial_i F_\kappa, i = 1,2$ evaluated at $(c_1^\star, c_2^\star)$. Given (c), we have proved that $(\hat{c}_1, \hat{c}_2, \hat{s})$ will be bounded for sufficiently large n, and therefore, uniform deviation bounds for $|\hat{F}_\kappa - F_\kappa|$ can also be established.

A critical consequence of our uniform deviation results is this: any optimizer of (4.9) possesses finite L_2-norm. Then, an adaptation of (Montanari et al., 2019, Section E) proves Theorem 3.2.

Step 4: Fixed point equations and final step. By standard analysis arguments (see Appendix A.3), the KKT conditions for the optimization problem (4.10) can be expressed as

$$\begin{aligned} &\Pi_{W^\perp}(G) + \psi^{-1/2}\Big[\partial_1 F_\kappa(c_1,c_2)W + c_2^{-1}\partial_2 F_\kappa(c_1,c_2)(\Lambda^{1/2}h - c_1 W)\Big] + s\cdot\Lambda^{-1/2}\partial\|h\|_{L^1(\mathcal{Q})} = 0, \\ &\text{and} \quad \|h\|_{L_1(\mathcal{Q})} = 1, \quad \text{where} \quad c_1 := \langle \Lambda^{1/2}h, W\rangle_{L_2(\mathcal{Q})}, \quad c_2 := \|\Pi_{W^\perp}(\Lambda^{1/2}h)\|_{L_2(\mathcal{Q})}. \end{aligned} \tag{4.12}$$

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

$$h = \frac{\mathbf{prox}_s\Big(\Lambda^{1/2}G + \psi^{-1/2}[\partial_1 F_\kappa(c_1,c_2) - c_1 c_2^{-1}\partial_2 F_\kappa(c_1,c_2)]\Lambda^{1/2}W\Big)}{\psi^{-1/2}c_2^{-1}\partial_2 F_\kappa(c_1,c_2)}. \tag{4.13}$$

Plugging this in the three equations displayed in (4.12), leads to the "fp equations … for (4.10)", referred to in Step 3, which is the exact same as the equation system (2.5), thus explaining the origin of the system. A similar analysis for (4.9) leads to the "empirical fp equations" referred to in Step 3 (see (A.11) for the specific form).

Finally, Corollary A.1 shows that $\tilde{\xi}^{(\infty,\infty)}_{\psi,\kappa}(\Lambda, W, G) = T(\psi,\kappa)$; together with (4.2) and (4.11), this completes the proof.

4.2 Proofs of Theorems 3.3 and 3.4

(Zhang and Yu, 2005) employs a re-scaling technique to establish that Boosting with infinitesimal stepsize agrees with the *min-L_1-norm* direction asymptotically. Since we care about the actual number of iterations in the Boosting algorithm (which translates to the number of selected features), here we provide a simple analysis of Boosting as a special instance of Mirror Descent, in conjunction with the re-scaling technique (Zhang and Yu, 2005) and the shrinkage technique (Telgarsky, 2013). Our analysis provides a sharp upper bound on the number of iterations of the algorithm, and is similar in spirit to (Collins et al., 2002), but with different executions.

Proposition 4.1 (Boosting as Mirror Descent)**.** *Consider the* Boosting Algorithm *stated in Section 2 Eqn. 2.12. Assume that $|X_{ij}| \le M$ 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*

$$T \ge \frac{2M^2}{\kappa^2_{n,\ell_1}} \log \frac{ne}{\epsilon}, \tag{4.14}$$

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

Corollary 4.1 (Boosting Converges to Max-L_1 Margin Direction)**.** *Consider the general Boosting algorithm with learning rate $\alpha_t(\beta) := \beta \cdot \eta_t^\top Z v_{t+1}$, where $\beta < 1$. Assume that $|X_{ij}| \le M$ for $i \in [n], j \in [p]$. Then after T iterations, the Boosting iterates θ_T converge to the Max-L_1 Margin Direction in the following sense: for any $0 < \epsilon < 1$,*

$$\kappa_{n,\ell_1} \ge \min_{i\in[n]} \frac{y_i x_i^\top \theta_T}{\|\theta_T\|_1} > \kappa_{n,\ell_1} \cdot (1-\epsilon), \tag{4.15}$$

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

To obtain Theorems 3.3 and 3.4, we choose $M(\delta) = \sqrt{(3+\delta)\log(np)}$ for arbitrarily small $\delta > 0$. Now, the entries X_{ij} are uniformly bounded above by M asymptotically almost surely, since $\mathbb{P}(\sup_{i\in[n],j\in[p]} |X_{ij}| \le M(\delta)) \lesssim np\exp(-M^2/2) = n^{-1-\delta}$ and $\sum_n n^{-1-\delta} < \infty$. Plugging in $\epsilon = 0.99$ in Proposition 4.1, with the aforementioned M, establishes the almost sure result in Theorem 3.4. The constant 12 can be justified since $\lim_{\delta\to 0} 2M^2(\delta)/\log n = 12$.

References

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.

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

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

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

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

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

Emmanuel J Candès and Pragya Sur. The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression. *arXiv preprint arXiv:1804.09753*, 2018.

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

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.

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

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

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

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

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

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.

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.

Tengyuan Liang and Alexander Rakhlin. Just interpolate: Kernel" ridgeless" regression can generalize. *The Annals of Statistics, to appear*, 2019.

Tengyuan Liang, Alexander Rakhlin, and Xiyu Zhai. On the risk of minimum-norm interpolants and restricted lower isometry of kernels. *arXiv preprint arXiv:1908.10292*, 2019.

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

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

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.

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.

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

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.

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

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

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.

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.

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

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

A Technical Lemmas

A.1 The Convex Gaussian Min-Max Theorem

For the convenience of the readers, we state Gordon's comparison inequality (Gordon, 1988) below (Thrampoulidis et al., 2015, Theorem 4). We state the form mentioned in (Montanari et al., 2019, Theorem 2).

Theorem A.1. *Let $C_1, C_2 \subset \mathbb{R}^n$ be two compact sets and let $R : C_1 \times C_2 \to \mathbb{R}$ be a continuous function. Let $X = (X_{i,j}) \in \mathbb{R}^{n\times p}, g \sim \mathcal{N}(0, I_p)$ and $h \sim \mathcal{N}(0, I_p)$ be independent vetors and matrices. Define*

$$Q_1(X) = \min_{w_1\in C_1} \max_{w\in C_2} w_1^\top X w_2 + R(w_1, w_2)$$
$$Q_2(g,h) = \min_{w_1\in C_1} \min_{w_1\in C_1} \max_{w_2\in C_2} \|w_2\| g^\top w_1 + \|w_1\| h^\top w_2 + R(w_1, w_2).$$

Then

1. *For all $t \in \mathbb{R}$,*
$$\mathbb{P}(Q_1(X) \le t) \leqq 2\mathbb{P}(Q_2(g,h) \le t).$$
2. *Suppose C_1 and C_2 are both convex, and R is convex concave in (w_1, w_2). Then, for all $t \in \mathbb{R}$,*
$$\mathbb{P}(Q_1(X) \ge t) \le 2\mathbb{P}(Q_2(g,h) \ge t).$$

A.2 Probability Results

Let $g \in \mathbb{R}^n$ be such that $g_i \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$. Recall the definitions of λ_j, w_j from Assumption 1 and (4.3) respectively, and denote the empirical distribution of $\{(\lambda_j, \sqrt{p}w_j, g_j)\}_{i=1}^n$ by $\mathcal{Q}_p$, that is,

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

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

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

where $F_\kappa(\cdot,\cdot)$ is given by (2.4).

Then from Proposition 2.1, we immediately obtain the following.

Lemma A.1. *Given any (ψ,κ) such that $\psi > \psi^{\downarrow}(\kappa)$, denote $(c_1^\star, c_2^\star, s^\star) \in \mathbb{R}\times\mathbb{R}_{>0}\times\mathbb{R}_{>0}$ to be the unique solution to the system (2.5). If a triplet $(c_1,c_2,s)\in\mathbb{R}\times\mathbb{R}_{>0}\times\mathbb{R}_{>0}$ satisfies for every $\epsilon>0$, there exists $\delta(\epsilon)>0$ small enough such that*

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

then, (c_1,c_2,s) must be ϵ-close to $(c_1^\star,c_2^\star),c_3^\star)$, that is,

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

We next turn to define different empirical versions of (A.2), which will be used later. To this end, recall that (4.8)

$$\hat{F}_\kappa(c_1,c_2) := \Big(\widehat{\mathbf{E}}_n[(\kappa - c_1 YZ_1 - c_2 Z_2)_+^2]\Big)^{1/2}, \tag{A.5}$$

and define

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

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

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

We are now in position to establish (4.11). Recall the finite n,p optimization problem

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

and the corresponding infinite-dimensional optimization problem given by

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

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

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

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

Proof of Proposition A.1. To begin with, recall (A.62)-(A.64)—together these establish the following fixed point equations

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

Note that the objective function in (A.8) is not convex in θ. Nonetheless, for any θ that minimizes the objective, the KKT conditions still hold. Thus, by arguments similar to that in the proof of Proposition 2.1, with $\theta/\sqrt{p}, \mathcal{Q}_p, \widehat{F}_k$ replacing $h,\mathcal{Q}_\infty,F_k$, we obtain the finite sample versions

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

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

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

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

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

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

Now it suffices to provide a uniform deviation bound for

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

in the region $c_1\in[0,M], c_2>0, s>0$ (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

$$\hat{C}^{\uparrow} := \psi^{-1/2}[\partial_1\hat{F}_\kappa(c_1,c_2) - c_1c_2^{-1}\partial_2\hat{F}_\kappa(c_1,c_2)] \tag{A.16}$$

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

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

$$Eqn.\ (A.15) \leq \tag{A.18}$$

$$(c_2 \vee 1)^{-1} \left\{ \frac{\|\Lambda^{1/2}G\|_{L_2(\mathcal{Q}_p)}\|\Lambda^{1/2}W\|_{L_2(\mathcal{Q}_p)} + |\hat{C}^{\uparrow}|\|\Lambda^{1/2}W\|^2_{L_2(\mathcal{Q}_p)}}{|\hat{C}^{\downarrow}C^{\downarrow}|}|\hat{C}^{\downarrow} - C^{\downarrow}| + \frac{\|\Lambda^{1/2}W\|^2_{L_2(\mathcal{Q}_p)}}{|C^{\downarrow}|}|\hat{C}^{\uparrow} - C^{\uparrow}| \right\} \tag{A.19}$$

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

$$Eqn.\ (A.15) \leq \frac{c_2 L + |c_2\hat{C}^{\uparrow}|L}{|c_2\hat{C}^{\downarrow}||c_2 C^{\downarrow}|}|(c_2\hat{C}^{\downarrow}) - (c_2 C^{\downarrow})| + \frac{L}{|c_2 C^{\downarrow}|}|(c_2\hat{C}^{\uparrow}) - (c_2 C^{\uparrow})| \tag{A.20}$$

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

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

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

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

$$Eqn.\ (A.15) \leq L' \cdot \frac{\log n}{\sqrt{n}}, \tag{A.21}$$

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 A.2

$$Eqn.\ (A.15) \leq (c_2 \vee 1)^{-1}\left(c_2 \frac{L + |\hat{C}^{\uparrow}|L}{|(c_2\hat{C}^{\downarrow}||c_2 C^{\downarrow}|}|(c_2\hat{C}^{\downarrow}) - (c_2 C^{\downarrow})| + c_2\frac{L}{|c_2 C^{\downarrow}|}|\hat{C}^{\uparrow} - C^{\uparrow}|\right) \tag{A.22}$$

$$\leq \frac{L + |\hat{C}^{\uparrow}|L}{|(c_2\hat{C}^{\downarrow}||c_2 C^{\downarrow}|}|(c_2\hat{C}^{\downarrow}) - (c_2 C^{\downarrow})| + \frac{L}{|c_2 C^{\downarrow}|}|\hat{C}^{\uparrow} - C^{\uparrow}| \tag{A.23}$$

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

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

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

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

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

$$\sup_{|c_1| \leq M, c_2 > 0, s > 0} (c_2 \vee 1)^{-1}|V_1^{(n,p)}(c_1, c_2, s) - V_1^{(\infty,p)}(c_1, c_2, s)| \leq \frac{\log n}{\sqrt{n}} \tag{A.24}$$

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

$$(c_2 \vee 1)^{-1}|V_1^{(\infty,p)}(c_1, c_2, s) - V_1^{(\infty,\infty)}(c_1, c_2, s)| \tag{A.25}$$

$$= \left|\left(\mathop{\mathbf{E}}_{(\Lambda,W,G)\sim\mathcal{Q}_p} - \mathop{\mathbf{E}}_{(\Lambda,W,G)\sim\mathcal{Q}_\infty}\right)(c_2 \vee 1)^{-1} f_1(\Lambda, W, G)\right|, \tag{A.26}$$

where

$$f_1(\Lambda, W, G) := \begin{pmatrix} \Lambda^{1/2} W \cdot \mathbf{prox}_s\left(\Lambda^{1/2}\Pi_{W^\perp}(G) + \psi^{-1/2}[\partial_1 F_\kappa(c_1,c_2) - c_1 c_2^{-1}\partial_2 F_\kappa(c_1,c_2)]\Lambda^{1/2} W\right) \\ \psi^{-1/2} c_2^{-1}\partial_2 F_\kappa(c_1,c_2) \end{pmatrix} \tag{A.27}$$

Since $\mathcal{Q}_p \overset{W_2}{\Longrightarrow} \mathcal{Q}_\infty$, for any function g that grows at most quadratically, (Ambrosio and Gigli, 2013, Proposition 2.4) states that

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

We now verify that f_1 satisfies the quadratic growth condition

$$|f_1(\Lambda,W,G)| \le \frac{|\Lambda W \Pi_{W^\perp}(G)| + |C^\uparrow||\Lambda W^2|}{|C^\downarrow|} \le \frac{G^2 + \Lambda^2 W^2 + |C^\uparrow|\cdot \Lambda W^2}{|C^\downarrow|} \ .$$

Further, for all $|c_1| \le M, 0 \le c_2 \le M, s \ge 0$, uniformly for Λ, W, G

$$\frac{(c_2 \vee 1)^{-1}|f_1(\Lambda,W,G)|}{1+\Lambda^2+W^2+G^2} \le \frac{c_2(G^2+\Lambda^2W^2) + |c_2 C^\uparrow|\Lambda W^2}{|c_2 C^\downarrow|(1+\Lambda^2+W^2+G^2)} < \infty \ , \tag{A.29}$$

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

$$\frac{(c_2 \vee 1)^{-1}|f_1(\Lambda,W,G)|}{1+\Lambda^2+W^2+G^2} \le \frac{(G^2+\Lambda^2W^2) + |C^\uparrow|\Lambda W^2}{|c_2 C^\downarrow|(1+\Lambda^2+W^2+G^2)} < \infty \ . \tag{A.30}$$

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

$$\lim_{n,p\to\infty} \sup_{|c_1|\le M, c_2>0, s>0} (c_2 \vee 1)^{-1}|V_1^{(\infty,p)}(c_1,c_2,s) - V_1^{(\infty,\infty)}(c_1,c_2,s)| = 0, \tag{A.31}$$

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

$$\lim_{n,p\to\infty} \sup_{|c_1|\le M, c_2>0, s>0} (c_2 \vee 1)^{-1}|V_1^{(n,p)}(c_1,c_2,s) - V_1^{(\infty,p)}(c_1,c_2,s)| = 0 \ . \tag{A.32}$$

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

$$\lim_{p\to\infty, p/n(p)\to\psi} \sup_{|c_1|\le M, c_2>0, s>0} (c_2 \vee 1)^{-1}|V_1^{(n,p)}(c_1,c_2,s) - V_1^{(\infty,p)}(c_1,c_2,s)| = 0, \quad a.s. \tag{A.33}$$

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

$$|V_2^{(n,p)}(c_1,c_2,s) - V_2^{(\infty,\infty)}(c_1,c_2,s)| \le |V_2^{(n,p)}(c_1,c_2,s) - V_2^{(\infty,p)}(c_1,c_2,s)| + |V_2^{(\infty,p)}(c_1,c_2,s) - V_2^{(\infty,\infty)}(c_1,c_2,s)| \ . \tag{A.34}$$

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

$$(c_2 \vee 1)^{-2}|V_2^{(n,p)}(c_1,c_2,s) - V_2^{(\infty,p)}(c_1,c_2,s)| \tag{A.35}$$

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

$$\left.\times\left(\frac{|\Lambda^{1/2}\Pi_{W^\perp}(G)| + |\hat{C}^\uparrow||\Lambda^{1/2}W|}{|\hat{C}^\downarrow C^\downarrow|}|\hat{C}^\downarrow - C^\downarrow| + \frac{|\Lambda^{1/2}W|}{|C^\downarrow|}|\hat{C}^\uparrow - C^\uparrow|\right)\right\}. \tag{A.36}$$

Again we divide the range of c_2 into two parts, $(0, M]$ and (M, ∞). For the first part, uniformly over $(c_1, c_2) \in [-M, M] \times (0, M]$, Lemma A.2 shows that

$$|c_2\hat{C}^{\downarrow} - c_2 C^{\downarrow}|, |c_2\hat{C}^{\uparrow} - c_2 C^{\uparrow}| \lesssim \frac{\log n}{\sqrt{n}} \ . \tag{A.37}$$

For the second part, uniformly over $(c_1, c_2) \in [-M, M] \times (M, \infty)$, Lemma A.2 shows that

$$|c_2\hat{C}^{\downarrow} - c_2 C^{\downarrow}|, |\hat{C}^{\uparrow} - C^{\uparrow}| \lesssim \frac{\log n}{\sqrt{n}} \ . \tag{A.38}$$

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

$$\sup_{|c_1| \le M, c_2 > 0, s > 0} (c_2 \vee 1)^{-2} |V_2^{(n,p)}(c_1, c_2, s) - V_2^{(\infty,p)}(c_1, c_2, s)| \le L' \cdot \frac{\log n}{\sqrt{n}} \ . \tag{A.39}$$

For the term,

$$(c_2 \vee 1)^{-2} |V_2^{(\infty,p)}(c_1, c_2, s) - V_2^{(\infty,\infty)}(c_1, c_2, s)| \tag{A.40}$$

$$\le (c_2 \vee 1)^{-2} \left| \left(\mathop{\mathbf{E}}_{(\Lambda,W,G)\sim\mathcal{Q}_p} - \mathop{\mathbf{E}}_{(\Lambda,W,G)\sim\mathcal{Q}_\infty} \right) (c_2 \vee 1)^{-1} f_2(\Lambda, W, G) \right|, \tag{A.41}$$

with

$$f_2(\Lambda, W, G) := \left(\frac{\mathbf{prox}_s\left(\Lambda^{1/2} \Pi_{W^\perp}(G) + \psi^{-1/2} [\partial_1 F_\kappa(c_1, c_2) - c_1 c_2^{-1} \partial_2 F_\kappa(c_1, c_2)] \Lambda^{1/2} W \right)}{\psi^{-1/2} c_2^{-1}, \partial_2 F_\kappa(c_1, c_2)} \right)^2 \tag{A.42}$$

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

$$\frac{(c_2 \vee 1)^{-2} |f_2(\Lambda, W, G)|}{1 + \Lambda^2 + W^2 + G^2} < \infty \ . \tag{A.43}$$

Therefore,

$$\lim_{n,p\to\infty} \sup_{|c_1| \le M, c_2 > 0, s > 0} (c_2 \vee 1)^{-2} |V_2^{(\infty,p)}(c_1, c_2, s) - V_2^{(\infty,\infty)}(c_1, c_2, s)| = 0 \tag{A.44}$$

$$\lim_{n,p\to\infty} \sup_{|c_1| \le M, c_2 > 0, s > 0} (c_2 \vee 1)^{-2} |V_2^{(n,p)}(c_1, c_2, s) - V_2^{(\infty,p)}(c_1, c_2, s)| = 0 \ . \tag{A.45}$$

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

$$\lim_{n,p\to\infty} \hat{c}_1 = c_1^\star, \quad \lim_{n,p\to\infty} \hat{c}_2 = c_2^\star, \quad \lim_{n,p\to\infty} \hat{s} = s^\star \tag{A.46}$$

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.

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

$$\|\hat{\theta}\|^2 \asymp \|\Lambda^{1/2}\hat{\theta}\|_2^2 = \langle w, \Lambda^{1/2}\hat{\theta}\rangle^2 + \|\Pi_{w^\perp}\hat{\theta}\|_2^2 = \hat{c}_1^2 + \hat{c}_2^2 \le 2(c_1^\star)^2 + 2(c_2^\star)^2 < 4R^2 \tag{A.47}$$

for some absolute constant $R > 0$, for sufficiently large n and p.

Given Eqn. A.46, one can verify by the KKT condition that the optimal value of finite n,p optimization problem $\hat{\xi}^{(n,p)}_{\psi,\kappa}(\lambda,w,g)$ can be expressed in the form

$$\hat{T}(\psi,\kappa;\hat{c}_1,\hat{c}_2,\hat{s}) := \psi^{-1/2}[\hat{F}_\kappa(\hat{c}_1,\hat{c}_2) - \hat{c}_1\partial_1\hat{F}_\kappa(\hat{c}_1,\hat{c}_2) - \hat{c}_2\partial_2\hat{F}_\kappa(\hat{c}_1,\hat{c}_2)] - \hat{s} \tag{A.48}$$

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

$$\lim_{n,p\to\infty} \sup_{c_1,c_2\in\mathcal{B}(c_1^\star,R),\mathcal{B}(c_2^\star,R)} |\hat{F}_\kappa(c_1,c_2) - F_\kappa(c_1,c_2)| = 0 \quad a.s. \tag{A.49}$$

By Lemma A.2, $\partial_1\hat{F}_\kappa$ and $\partial_2\hat{F}_\kappa$ all satisfy uniform convergence over $|c_1| \le M, c_2 > 0$. Therefore

$$\lim_{n,p\to\infty} \hat{T}(\psi,\kappa;\hat{c}_1,\hat{c}_2,\hat{s}) \tag{A.50}$$

$$= \lim_{n,p\to\infty} \psi^{-1/2}[F_\kappa(\hat{c}_1,\hat{c}_2) - \hat{c}_1\partial_1 F_\kappa(\hat{c}_1,\hat{c}_2) - \hat{c}_2\partial_2 F_\kappa(\hat{c}_1,\hat{c}_2)] - \hat{s} \quad \text{by uniform convergence} \tag{A.51}$$

$$= \psi^{-1/2}[F_\kappa(c_1^\star,c_2^\star) - c_1^\star\partial_1 F_\kappa(c_1^\star,c_2^\star) - c_2^\star\partial_2 F_\kappa(c_1^\star,c_2^\star)] - s^\star = T(\psi,\kappa)\ . \tag{A.52}$$

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

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

□

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

$$\sup_{|c_1|\le M, c_2>0} |\partial_i\hat{F}_\kappa(c_1,c_2) - \partial_i F_\kappa(c_1,c_2)| \le \frac{C\log n}{\sqrt{n}} \tag{A.53}$$

as long as c_1,c_2 are in the region with $F_\kappa(c_1,c_2) \ne 0$ (this can be proved since $\min_{|c_1|\le M,c_2>0} F_\kappa(c_1,c_2) > 0$).

Proof. The proof uses a key self-normalization property of the partial derivatives of F_κ, that ensure good concentration behavior even when c_2 is large. Note that

$$\partial_1\hat{F}_\kappa(c_1,c_2) = -\frac{\hat{\mathbf{E}}_n[YZ_1\sigma(\kappa - c_1YZ_1 - c_2Z_2)]}{(\hat{\mathbf{E}}_n[\sigma^2(\kappa - c_1YZ_1 - c_2Z_2)])^{1/2}}, \tag{A.54}$$

$$\partial_2\hat{F}_\kappa(c_1,c_2) = -\frac{\hat{\mathbf{E}}_n[Z_2\sigma(\kappa - c_1YZ_1 - c_2Z_2)]}{(\hat{\mathbf{E}}_n[\sigma^2(\kappa - c_1YZ_1 - c_2Z_2)])^{1/2}} \tag{A.55}$$

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

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

In the first region, where $(c_1, c_2) \in [-M, M] \times (0, M]$, it is easy to verify that $R_1(c_1, c_2) := YZ_1\sigma(\kappa - c_1 YZ_1 - c_2 Z_2)$, $R_2(c_1, c_2) := Z_2\sigma(\kappa - c_1 YZ_1 - c_2 Z_2)$ and $R_0(c_1, c_2) := \sigma^2(\kappa - c_1 YZ_1 - c_2 Z_2)$ are all sub-exponential random variables with at most a constant sub-exponential parameter. Then a simple ϵ-covering on the bounded region $[-M, M] \times (0, M]$ gives us that with probability at least $1 - n^{-2}$,

$$\sup_{(c_1,c_2)\in(0,M]\times(0,M]} |\hat{\mathbf{E}}_n[R_j(c_1, c_2)] - \mathbf{E}[R_j(c_1, c_2)]| \lesssim \frac{\log n}{\sqrt{n}}, \forall j \in 0, 1, 2 \ . \tag{A.56}$$

Recall that $\mathbf{E}[R_0(c_1, c_2)] = F_\kappa(c_1, c_2) \neq 0$. Then for n large enough, the claim follows since

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

w.p. at least $1 - n^{-2}$ 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 $\partial_i \hat{F}_\kappa(c_1, c_2)$

$$\partial_1 \hat{F}_\kappa(c_1, c_2) = -\frac{\hat{\mathbf{E}}_n[YZ_1\sigma(\kappa c_2^{-1} - c_1 c_2^{-1} YZ_1 - Z_2)]}{(\hat{\mathbf{E}}_n[\sigma^2(\kappa c_2^{-1} - c_1 c_2^{-1} YZ_1 - Z_2)])^{1/2}} \tag{A.57}$$

$$\partial_2 \hat{F}_\kappa(c_1, c_2) = -\frac{\hat{\mathbf{E}}_n[Z_2\sigma(\kappa c_2^{-1} - c_1 c_2^{-1} YZ_1 - Z_2)]}{(\hat{\mathbf{E}}_n[\sigma^2(\kappa c_2^{-1} - c_1 c_2^{-1} YZ_1 - Z_2)])^{1/2}} \tag{A.58}$$

Now the regions for the parameers of interest are bounded since

$$(c_2^{-1}, c_1 c_2^{-1}) \in [0, 1/M) \times (-1, 1). \tag{A.59}$$

Once again, $R_1'(c_1, c_2) := YZ_1\sigma(\kappa c_2^{-1} - c_1 c_2^{-1} YZ_1 - Z_2)$, $R_2'(c_1, c_2) := Z_2\sigma(\kappa c_2^{-1} - c_1 c_2^{-1} YZ_1 - Z_2)$ and $R_0(c_1, c_2) := \sigma^2(\kappa c_2^{-1} - c_1 c_2^{-1} YZ_1 - Z_2)$ are all sub-exponential random variables with sub-exponential parameter at most a constant. A standard ϵ-covering on $(c_2^{-1}, c_1 c_2^{-1})$ completes the proof for the region $(c_1, c_2) \in (0, M] \times (M, \infty)$. □

Proof of Theorem 3.2. The proof follows by an adaptation of (Montanari et al., 2019, Section E), utilizing the bounds (A.47).

□

A.3 Uniqueness Results

We next present the proofs of Propositions 2.1 and 2.2.

Proof of Proposition 2.1. To analyze the equation system (2.5), we will, in fact, begin by examining the objective function in (A.9) as a function of h, that is, define

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

and consider the optimization problem

$$\text{minimize} \qquad \mathcal{R}_{\psi,\kappa,Q_\infty}(h) \qquad \text{s.t.} \qquad \|h\|_{L_1(Q_\infty)} \leq 1. \tag{A.60}$$

By arguments similar to that in (Montanari et al., 2019, Section B.3.1), one can show that the function $h \to \mathcal{R}_{\psi,\kappa,Q_\infty}$ is strictly convex and, the minimum of the optimization problem (A.60) is achieved at a unique function $h^\star \in \mathcal{L}^2(Q_\infty)$. Then the unique minimizer is determined by the KKT conditions, which in this case can be expressed as

$$\Lambda^{1/2}\Pi_{W^\perp}(G)+\psi^{-1/2}\Lambda^{1/2}\left[\partial_1 F_\kappa(c_1,c_2)W+\partial_2 F_\kappa(c_1,c_2)\Pi_{W^\perp}(Z)\right]+s\cdot\partial\|h\|_{L_1(\mathcal{Q}_\infty)}=0,$$
$$s(1-\|h\|_{L_1(\mathcal{Q}_\infty)})=0, \tag{A.61}$$
$$s\geq 0, \|h\|_{L_1(\mathcal{Q}_\infty)}\leq 1.$$

Above, Z is given by

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

Now, if $\psi>\psi^{\downarrow}(\kappa)$, and Assumptions 1-3 are satisfied, the conditions B1-B3 in (Montanari et al., 2019, Lemma B.4) are satisfied with $\zeta=\left(\mathop{\mathbf{E}}_{(\Lambda,W)\sim\mu}|\Lambda^{-1/2}W|\right)^{-1}$. Note that this is different from the choice of ζ considered in (Montanari et al., 2019). With this choice, an adaptation of the arguments in (Montanari et al., 2019, Section B.3.3) with appropriate changes in the constants $M,\Delta,\gamma_+(\Delta),\gamma_-(\Delta),\tilde{\gamma}_+(\Delta)$ and $\tilde{\gamma}_-(\Delta)$ yields that for any minimizer h and the corresponding dual variable s, we have $s>0$ and $\|\Pi_{W^\perp}(\Lambda^{1/2}h)\|>0$. Denoting $c_1:=\langle\Lambda^{1/2}h,W\rangle_{L_2(\mathcal{Q})}$ and $c_2:=\|\Pi_{W^\perp}(\Lambda^{1/2}h)\|_{L_2(\mathcal{Q})}$, the KKT conditions can then be rewritten as

$$\Pi_{W^\perp}(G)+\psi^{-1/2}\left[\partial_1 F_\kappa(c_1,c_2)W+c_2^{-1}\partial_2 F_\kappa(c_1,c_2)(\Lambda^{1/2}h-c_1W)\right]+s\cdot\Lambda^{-1/2}\partial\|h\|_{L^1(\mathcal{Q}_\infty)}=0$$
$$\|h\|_{L_1(\mathcal{Q}_\infty)}=1. \tag{A.62}$$

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

$$h^\star=\frac{\mathbf{prox}_s\left(\Lambda^{1/2}G+\psi^{-1/2}[\partial_1 F_\kappa(c_1,c_2)-c_1c_2^{-1}\partial_2 F_\kappa(c_1,c_2)]\Lambda^{1/2}W\right)}{\psi^{-1/2}c_2^{-1}\partial_2 F_\kappa(c_1,c_2)}. \tag{A.63}$$

Plugging this in the system

$$c_1=\langle\Lambda^{1/2}h^\star,W\rangle_{L_2(\mathcal{Q}_\infty)},\qquad c_1^2+c_2^2=\|\Lambda^{1/2}h^\star\|^2_{L_2(\mathcal{Q}_\infty)},\qquad \|h^\star\|_{L_1(\mathcal{Q}_\infty)}=1 \tag{A.64}$$

yields the fixed point equations (2.5). 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.62) are also unique and, furthermore, c_2 and s are strictly positive. □

Proof of Proposition 2.2. The proof follows from to (Montanari et al., 2019, Section B.5) and, is therefore, omitted. □

We obtain a key representation for $\tilde{\xi}^{(\infty,\infty)}_{\psi,\kappa}(\Lambda,W,G)$ as a byproduct of the above. On taking inner products with $\Lambda^{1/2}h$ on both sides of the first equation in (A.62) leads to the following.

Corollary A.1. *Under the assumptions of Proposition 2.1, the minimum value of the optimization problem (A.60) is given by*

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

where $(c_1,c_2,s)\in\mathbb{R}\times\mathbb{R}_{>0}\times\mathbb{R}_{>0}$ forms the unique solution to (2.5). Hence, the above equals $T(\psi,\kappa)$ defined in (2.11).

A.4 Optimization Results

Proof of Proposition 4.1. We prove the convergence of *Boosting Algorithm* as a special case of Mirror Descent. In fact, we will prove the result for two scenarios: (1) AdaBoost, with $X_{ij} \in \{\pm 1\}$, (2) *Boosting Algorithm* in Eqn. 2.12, with $|X_{ij}| \leq M$ and a shrinkage on the learning rate. For $x \in \mathbb{R}^n$, define the entropy

$$R(x) = \sum_{i=1}^{n} x[i] \log(x[i]) + I_{\Delta_n}(x) \ . \tag{A.65}$$

The Fenchel conjugate of R, denoted by $R^\star$, reads,

$$R^\star(x) = \log\left(\sum_{i=1}^{n} \exp(x[i])\right) \ . \tag{A.66}$$

It is clear that R is 1-strongly convex w.r.t. the L_1 norm, and that $R^\star$ is 1-strongly smooth w.r.t. the L_∞ norm.

First, let us recall the dual formulation of L_1 margin

$$\kappa_{n,\ell_1} = \max_{\|\theta\|_1 \leq 1} \min_{j \in [p]} e_j^\top Z\theta = \min_{\eta \in \Delta_n} \max_{\|\theta\|_1 \leq 1} \eta^\top Z\theta = \min_{\eta \in \Delta_n} \|Z^\top \eta\|_\infty \tag{A.67}$$

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.

- L_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 $\alpha_t = \frac{1}{2} \log \frac{1+\gamma_t}{1-\gamma_t}$. One can see this since

$$\min_{v \in \{\pm e_j\}_{j \in [p]}} \sum_{i \in [n]} \eta_t[i] \cdot I_{y_i x_i^\top v \leq 0} = \min_{v \in \{\pm e_j\}_{j \in [p]}} \sum_{i \in [n]} \eta_t[i] \cdot I_{-y_i x_i^\top v \geq 0} \tag{A.68}$$

$$= \frac{1}{2}(-\max_{j \in [p]} |\eta_t^\top Z e_j| + 1). \tag{A.69}$$

- Updates on $\eta \in \Delta_n$ (Mirror Descent):

$$\nabla R(\eta_t) = -Z\theta_t \tag{A.70}$$

$$Z\theta_{t+1} = Z\theta_t + \alpha_t Z v_{t+1} \ . \tag{A.71}$$

$$\nabla R^\star(-Z\theta_{t+1}) = \eta_{t+1} \tag{A.72}$$

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

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

For the (2) *Boosting Algorithm* in Eqn. 2.12, with $|X_{ij}| \leq M$. Define

$$\alpha_t(\beta) = \beta \cdot \eta_t^\top Z v_{t+1} \tag{A.73}$$

then

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

with $\beta = 1/M^2$

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

$$
R^\star(-Z\theta_T) - R^\star(-Z\theta_0) \le -\frac{\sum_{t=0}^{T-1}\gamma_t^2}{2M^2} \le -T\frac{\kappa_{n,\ell_1}^2}{2M^2} \tag{A.74}
$$

$$
\sum_{i\in[n]} I_{-y_i x_i^\top \theta_T > 0} \le \sum_{i\in[n]} \exp(-y_i x_i^\top \theta_T) = \exp(R^\star(-Z\theta_T)) \le ne\exp(-T\frac{\kappa_{n,\ell_1}^2}{2M^2}) \ . \tag{A.75}
$$

The proof now is complete.

□

Proof of Corollary 4.1. The proof follows from Proposition 4.1 and a re-scaling technique in Zhang and Yu (2005)'s asymptotic analysis. Here instead, we give a non-asymptotic result. For any $\kappa > 0$

$$
\sum_{i\in[n]} I_{\frac{y_i x_i^\top \theta_t}{\|\theta_t\|_1} \le \kappa} \le \sum_{i\in[n]} \exp(\kappa\|\theta_t\|_1 - y_i x_i^T \theta_t) \tag{A.76}
$$

$$
\le \exp(\kappa\|\theta_t\|_1)\exp\left(R^\star(-Z\theta_t)\right) \tag{A.77}
$$

with $R^\star$ defined in (A.66). Due to the proof in Proposition 4.1, we know

$$
R^\star(-Z\theta_T) \le R^\star(-Z\theta_0) - \sum_{t=0}^{T-1}\left(\beta\gamma_t^2 - \frac{\beta^2\gamma_t^2}{2}M^2\right) \tag{A.78}
$$

$$
\le \log(ne) - \sum_{t=0}^{T-1}\beta\gamma_t\left[\gamma_t - \frac{\beta}{2}\gamma_t M^2\right] \ . \tag{A.79}
$$

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

$$
\|\theta_T\| \le \sum_{t=0}^{T-1}\|\alpha_t v_{t+1}\| \le \sum_{t=0}^{T-1}\beta\gamma_t \ . \tag{A.80}
$$

Therefore

$$
\sum_{i\in[n]} I_{\frac{y_i x_i^\top \theta_t}{\|\theta_t\|_1} \le \kappa} \le ne\cdot\exp\left\{-\sum_{t=0}^{T-1}\beta\gamma_t\left[\gamma_t - \frac{\beta}{2}\gamma_t M^2 - \kappa\right]\right\} \tag{A.81}
$$

Recall that $\gamma_t \ge \kappa_{n,\ell_1}^2$ for all t, we know that

$$
\sum_{i\in[n]} I_{\frac{y_i x_i^\top \theta_t}{\|\theta_t\|_1} \le \kappa} \le ne\cdot\exp\left(-T\beta\kappa_{n,\ell_1}^2\left[\kappa_{n,\ell_1}^2(1 - \frac{\beta M^2}{2}) - \kappa\right]\right) \ . \tag{A.82}
$$

Choose

$$\beta = \frac{1-\kappa/\kappa_{n,\ell_1}^2}{M^2} \tag{A.83}$$

$$T \geq \log(1.01ne) \cdot \frac{2M^2\kappa_{n,\ell_1}^{-2}}{(1-\kappa/\kappa_{n,\ell_1})^2}, \tag{A.84}$$

we know

$$\sum_{i\in[n]} I_{\frac{y_i x_i^\top \theta_t}{\|\theta_t\|_1} \leq \kappa} \leq \frac{1}{1.01} < 1 \tag{A.85}$$

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

$$T = \log(1.01ne) \cdot \frac{2M^2\kappa_{n,\ell_1}^{-2}}{\epsilon^2} \tag{A.86}$$

we must have that

$$\min_{i\in[n]} \frac{y_i x_i^\top \theta_T}{\|\theta_T\|_1} > \kappa_{n,\ell_1} \cdot (1-\epsilon). \tag{A.87}$$

□

For completeness, we show that the min-L_1-norm interpolation, is equivalent to the max-L_1-margin solution. We use this fact several places in the main text.

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

$$\textit{Formulation I:} \quad I^\star := \max\left\{\kappa \mid \exists\theta,\ \|\theta\|_1 \leq 1,\ \ s.t.\ \ \forall i \leq n,\ y_i x_i^\top \theta \geq \kappa\right\} \tag{A.88}$$

$$\textit{Formulation II:} \quad II^\star := \min\ \|\theta\|_1,\ \ s.t.\ \ \forall i \leq n,\ y_i x_i^\top \theta \geq 1 \tag{A.89}$$

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

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

□