# The Performance Analysis of Generalized Margin Maximizer (GMM) on Separable Data

**Fariborz Salehi, Ehsan Abbasi, Babak Hassibi** [1]

## Abstract

Logistic models are commonly used for binary classification tasks. The success of such models has often been attributed to their connection to maximum-likelihood estimators. It has been shown that gradient descent algorithm, when applied on the logistic loss, converges to the max-margin classifier (a.k.a. hard-margin SVM). The performance of the max-margin classifier has been recently analyzed in (Montanari et al., 2019; Deng et al., 2019). Inspired by these results, in this paper, we present and study a more general setting, where the underlying parameters of the logistic model possess certain structures (sparse, block-sparse, low-rank, etc.) and introduce a more general framework (which is referred to as "Generalized Margin Maximizer", GMM). While classical max-margin classifiers minimize the 2-norm of the parameter vector subject to linearly separating the data, GMM minimizes any arbitrary convex function of the parameter vector. We provide a precise analysis of the performance of GMM via the solution of a system of nonlinear equations. We also provide a detailed study for three special cases: (1) $\ell_2$-GMM that is the max-margin classifier, (2) $\ell_1$-GMM which encourages sparsity, and (3) $\ell_\infty$-GMM which is often used when the parameter vector has binary entries. Our theoretical results are validated by extensive simulation results across a range of parameter values, problem instances, and model structures.

## 1. Introduction

Machine learning models have been very successful in many applications, ranging from spam detection, face and pattern recognition, to the analysis of genome sequencing and financial markets. However, despite this indisputable success, our knowledge on why the various machine learning methods exhibit the performances they do is still at a very early stage. To make this gap between the theory and the practice narrower, researchers have recently begun to revisit simple machine learning models with the hope that understanding their performance will lead the way to understanding the performance of more complex machine learning methods.
More specifically, studies on the performance of diffrent classifiers for binary classification dates back to the seminal work of Vapnik in the 1980's (Vapnik, 1982). In an effort to find the "optimal" hyperplane that separates the data, he presented an upper bound on the test error which is inversely proportional to the margin (minimum distance of the datapoints to the separating hyperplane), and concluded that the max-margin classifier is indeed the desired classifier. It has also been observed that to construct such optimal hyperplanes one only has to take into acconnt a small amount of the training data, the so-called support vectors (Cortes & Vapnik, 1995).
In this paper, we challenge the conventional wisdom by showing that when the underlying parameter has certain structure one can come up with classifiers that outperform the max-margin classifier. We introduce the **G**eneralized **M**argin **M**aximizer (**GMM**) which takes into account the structure of the underlying parameter as well as the minimimum distance of the datapoints to the separating hyperplane. We provide sharp asymptotic results on various performance measures (such as the generalization error) of GMM and show that an appropriate choice of the potential function can in fact improve the resulting estimator.

### 1.1. Prior work

There have been many recent attempts to understand the generalization behavior of simple machine learning models (Bartlett et al., 2019; Mei & Montanari, 2019; Xu & Hsu, 2019; Belkin et al., 2018; Hastie et al., 2019). Most of these studies focus on the least-squares/ridge regression, where the loss function is the squarred $\ell_2$-norm, and derive sharp asymptotics on the performance of the estimator. In particular, in (Hastie et al., 2019; Kini & Thrampoulidis, 2020) the authors have shown that the minimum-norm least

[1]Department of Electrical Engineering, California Institute of Technology, Pasadena, California, USA. Correspondence to: Fariborz Salehi <fsalehi@caltech.com>.

*Proceedings of the $37^{th}$ International Conference on Machine Learning*, Online, PMLR 119, 2020. 

square solution demonstrates the so-called "double-descent" behavior (Belkin et al., 2019).

A more recent line of research studies the generalization performance of gradient descent (GD) for binary classification. It has been shown (Soudry et al., 2018)) that for a separable dataset, GD (when applied on the logistic loss) converges in direction to the max-margin classifier (a.k.a. hard-margin SVM). The performance of max-margin classifier has been recently analyzed in two independent works (Montanari et al., 2019; Deng et al., 2019).

### 1.2. Summary of contributions

Inspired by the recent results in understanding the performance of the max-margin classifier, in this paper we introduce and study a more general framework. We assume the underlying parameters possess certain structure (e.g. sparse) and introduce the generalized margin maximizer (GMM) as the solution of a convex optimization problem whose objective function encourages the structure.

We analyze the performance of GMM in the high-dimensional regime where both the number of parameters, $p$, and the number of samples $n$ grows, and analyze the asymptotic performance as a function of the overparameterization ratio $\delta := \frac{p}{n} > 0$. First, we provide the phase transition condition for the separability of data (i.e., derive the exact value of $\delta^*$ such that the data is separable for all $\delta > \delta^*$[1].) Consequently, we analyze the performance in the interpolating regime ($\delta > \delta^*$). To the best of our knowledge, this is the first theoretical result that provides sharp asymptotics on the performance of GMM classifiers on separable data. For our analysis, we exploit the **C**onvex **G**aussian **M**in-max **T**heorem (**CGMT**) (Stojnic, 2013; Thrampoulidis et al., 2015) which is a strengthened version of a classical Gaussian comparison inequality due to Gordon (Gordon, 1985). This framework replaces the original optimization with another optimization problem that has a similar performance, yet is much simpler to analyze as it becomes nearly separable. Previously, the CGMT has been successfully applied to derive the precise performance in a number of applications such as regularized M-estimators (Thrampoulidis et al., 2018), analysis of the generalized lasso (Miolane & Montanari, 2018; Thrampoulidis et al., 2015), data detection in massive MIMO (Abbasi et al., 2019; Atitallah et al., 2017; Thrampoulidis et al., 2019), and PhaseMax in phase retrieval (Dhifallah et al., 2018; Salehi et al., 2018a;b).

More recently, this framework has been employed in a series of works by multiple groups of researchers to characterize the performance of the logistic loss minimizer in binary classification (Salehi et al., 2019; Taheri et al., 2019). Furthermore, in an analogous avenue of research, the CGMT framework has been utilized to study the generalization behavior of the gradient descent algorithm in the interpolating regime, where there exists a (nonempty) set of parameters that perfectly fit the training data (Montanari et al., 2019; Deng et al., 2019).

The organization of the paper is as follows: In Section 2 we mathematically introduce the problem and the notations used in the paper. Section 3 contains the main results of the paper where we first provide the asymptotic phase transition on the separability of the data, and then in our main theorem, we present the precise performance analysis of GMM, which then be used to compute the generalization error. We investigate our theoretical findings for three specific cases of potential functions in Section 4. Numerical simulations for the genralization error of the GMM classifiers are presented in Section 5. We should note that most technical derivations of the results presented in the paper are deferred to the Appendix.

## 2. Preliminaries

### 2.1. Notations

Here, we gather the basic notations that are used throughout the paper. $X \sim p_X$ denotes that the random variable $X$ has a density $p_X$. $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the multivariate Gaussian distribution with mean $\boldsymbol{\mu}$, and covariance $\boldsymbol{\Sigma}$, and $\textsc{Rad}(p)$, for $p \in [0,1]$, is the symmetric bernouli random variable which takes the value $+1$ with ptobability $p$, and $-1$ with probability $1-p$. $\xrightarrow{D}$, and $\xrightarrow{P}$ represent convergence in distribution and in probability, respectively. Bold lower letters are reserved for vectors, and upper letters are for matrices. $\mathbf{1}_d$, and $\mathbf{I}_d$ respectively represent the all-one vector and the identity matrix in dimension $d$. For a vector $\mathbf{v}$, $v_i$ denotes its $i$-th entry, and $\|\mathbf{v}\|_p$ (for $p \geq 1$), is its $\ell_p$ norm, where we remove the subscript when $p = 2$. For a scalar $t \in \mathbb{R}$, $(t)_+ = \max(t, 0)$ denotes its positive part, and $\textsc{Sign}(t)$ indicates its sign.

A function $f : \mathbb{R}^d \to \mathbb{R}$ is called (invariantly) separable, when for all $\mathbf{w} \in \mathbb{R}^d$, $f(\mathbf{w}) = \sum_{i=1}^d \tilde{f}(w_i)$, for a real-valued function $\tilde{f}$. For a function $\Phi : \mathbb{R}^d \to \mathbb{R}$, the Moreau envelope associated with $\Phi(\cdot)$ is defined as,

$$M_\Phi(\mathbf{v}, t) = \min_{\mathbf{x} \in \mathbb{R}^d} \ \frac{1}{2t}||\mathbf{v} - \mathbf{x}||^2 + \Phi(\mathbf{x}) \,, \tag{1}$$

and the proximal operator is the solution to this optimization, i.e.,

$$\text{Prox}_{t\Phi(\cdot)}(\mathbf{v}) = \arg\min_{\mathbf{x} \in \mathbb{R}^d} \ \frac{1}{2t}||\mathbf{v} - \mathbf{x}||^2 + \Phi(\mathbf{x}) \,. \tag{2}$$

Finally, the function $\Phi(\cdot)$ is said to be locally-Lipschitz if for any $M > 0$, there exists a constant $L_M$, such that,

$$\forall \mathbf{u}, \mathbf{v} \in [-M, +M]^d, \ \ |\Phi(\mathbf{u}) - \Phi(\mathbf{v})| \leq L_M \left\|\mathbf{u} - \mathbf{v}\right\| . \tag{3}$$

[1] Concurrent to the submission of this paper, a similar phase transition has been demonstrated in (Kini & Thrampoulidis, 2020) for a somewhat different model.

### 2.2. Mathematical setup

We consider the problem of binary classification, having a set of training data, $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, where each of the sample points consists of a $p$-dimensional feature vector, $\mathbf{x}_i$, and a binary label, $y_i \in \{\pm 1\}$. We assume that the dataset $\mathcal{D}$ is generated from a logistic-type model with the underlying parameter $\mathbf{w}^\star \in \mathbb{R}^p$. This means that

$$y_i \sim \textsc{Rad}(\rho(\mathbf{x}_i^T \mathbf{w}^\star)) \,, \quad i = 1, \dots, n \,, \tag{4}$$

where $\rho : \mathbb{R} \to [0, 1]$ is a non-decreasing function and is often referred to as the link function. A commonly-used instance of the link function is the standard logistic function defined as $\rho(t) := \frac{1}{1+e^{-t}}$.

When $n/p$ is sufficiently large, i.e., when we have access to a sufficiently large number of samples, the maximum-likelihood estimator( $\hat{\mathbf{w}}_{ML}$) is well-defined. In such settings, the MLE is often the estimator of choice due to its desirable properties in the classical statistics. Sur and Candès (Sur & Candès, 2018) have recently studied the performance of the MLE in logistic regression in the high-dimensional regime, where the number of observations and parameters are comparable, and show, among other things, that the maximum likelihood estimator is biased. Their results have been extended to regularized logistic regression (Salehi et al., 2019), assuming some prior knowledge on the structure of the data. In particular, it has been observed that, when the regularization parameter is tuned properly, the regularized logistic regression can outperform the MLE.

Inspired by the recent results on analyzing the generalization error of machine learning models, in this paper, we study the generalization error of binary classification, in a regime of parameters known as the interpolating regime. Here, the assumption is that there exists a parameter vector that can perfectly fit (interpolate) the data, i.e.,

$$\exists \mathbf{w}_0 \;\text{ s.t. }\; \textsc{Sign}(\mathbf{w}_0^T \mathbf{x}_i) = y_i, \;\; \text{for } i = 1, 2, \dots, n. \tag{5}$$

Let $\mathcal{W}$ denote the set of all the parameters that interpolate the data.

$$\mathcal{W} = \{\mathbf{w} \in \mathbb{R}^p : \textsc{Sign}(\mathbf{w}^T \mathbf{x}_i) = y_i \,, \;\; \text{for } 1 \le i \le n.\}. \tag{6}$$

It has been observed that in many machine learning tasks, the iterative solvers that minimize the loss function often converge to one of the points in the set $\mathcal{W}$ (the training error converges to zero). Therefore, one can (qualitatively) pose the following important (yet still mysterious) question:

> Which point(s) in $\mathcal{W}$ is (are) "better" estimator(s) of the actual parameter, $\mathbf{w}^\star$?

In an attempt to find an answer to this question, we focus on the simple (yet fundamental) model of binary clasification. We assume that the underlying parameter, $\mathbf{w}^\star$ possesses certain structure (sparse, low-rank, block-sparse, etc.), and consider a locally-Lipschitz and convex function $\psi : \mathbb{R}^p \to \mathbb{R}$ which encourages this structure. We introduce the *Generalized Margin Maximizer* (GMM) as the solution to the following optimization:

$$\begin{aligned} \min_{\mathbf{w} \in \mathbb{R}^p} \quad & \psi(\mathbf{w}) \\ \text{s.t. } & y_i(\mathbf{x}_i^T \mathbf{w}) \ge 1, \;\; \text{for } 1 \le i \le n. \end{aligned} \tag{7}$$

It is worth noting that the condition on the separability of the dataset is crucial for the optimization program (7) to have a feasible point.

**Remark 1.** *It can be shown that when $\psi(\cdot)$ is absolutely scalable*[2]*, the GMM can be found by solving the following equivalent optimization program,*

$$\max_{\mathbf{w} \in \mathbb{R}^d} \frac{\psi(\mathbf{w})}{\min\limits_{1 \le i \le n} y_i(\mathbf{x}_i^T \mathbf{w})} = \max_{\mathbf{w} \in \mathbb{R}^d} \frac{\|\mathbf{w}\|}{\min\limits_{1 \le i \le n} y_i(\mathbf{x}_i^T \mathbf{w})} \times \frac{\psi(\mathbf{w})}{\|\mathbf{w}\|}. \tag{8}$$

*The first multiplicative term on the right indicates the margin associated with the separator $\mathbf{w}$, and the second term, $\frac{\psi(\mathbf{w})}{\|\mathbf{w}\|}$ takes into account the structure of the model. Hence, we refer to the objective function in the optimization (8) as the generalized margin, and the solution to this optimization is called the generalized margin maximizer (GMM).*

In this paper, we study the linear asymptotic regime in which the problem dimensions $p,\ n$ grow to infinity at a proportional rate, $\delta := \frac{p}{n} > 0$. Our main result characterizes the performance of the solution of (7), $\hat{\mathbf{w}}$, in terms of the ratio, $\delta$, and the signal strength, $\kappa := \frac{\|\mathbf{w}^\star\|}{\sqrt{p}}$. We assume that the datapoints, $\{\mathbf{x}_i\}_{i=1}^n$, are drawn independently from the Gaussian distribution. Our main result characterizes the performance of the resulting estimator through the solution of a system of five nonlinear equations with five unknowns. In particular, as an application of our main result, we can accurately predict the generalization error of the resulting estimator.

## 3. Main Results

In this section, we present the main results of the paper, that is the characterization of the performance of the generalized margin maximizers. Our results are represented in terms of a summary functional, $c_t(\cdot, \cdot)$, which incorporates the informaiton about the underlying model.

**Definition 1.** *For the parameter $t > 0$, the function $c_t : \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}_+$ is defined as,*

$$c_t(s, r) = \mathbb{E}\left[(1 - tsZ_1 Y - rZ_2)_+^2\right], \tag{9}$$

*where $Z_1, Z_2 \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$, and $Y \sim \textsc{Rad}(\rho(tZ_1))$.*

[2] A function $f : \mathbb{R}^d \to \mathbb{R}$ is absolutely scalable when,

$$\forall \mathbf{v} \in \mathbb{R}^d, \forall \alpha \in \mathbb{R}, \;\; f(\alpha \mathbf{v}) = |\alpha| f(\mathbf{v}).$$

All $\ell_p$ norms, for example, are absolutely scalable.

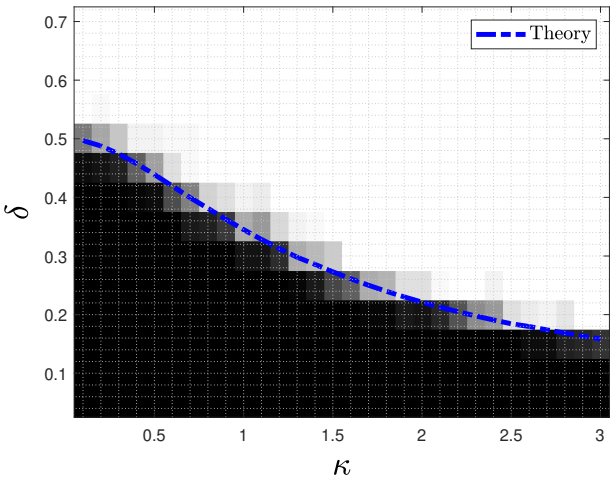


Figure 1: The phase transition, $\delta^*$, for the separability of the dataset, where the feature vector, $\mathbf{x}_i$ is drawn from the Gaussian distribution, $\mathcal{N}(\mathbf{0}, \frac{1}{p}\mathbf{I}_p)$, and the labels are $y_i \sim \text{RAD}\big(\rho(\mathbf{x}_i^T\mathbf{w}^\star)\big)$, for $\rho(z) = \frac{e^t}{e^t+e^{-t}}$. The empirical result is the average over 20 trials with $p = 150$, and the theoretical results are from Theorem 1.

### 3.1. Asymptotic phase transition

Here, we provide the necessary and sufficient condition for the separability of the data.

**Theorem 1** (Phase transition)**.** *Consider the generalized max margin optimization defined in Section 2.2. As $n, p \to \infty$ at a fixed overparameterization ratio $\delta := \frac{p}{n} \in (0, \infty)$, this optimization program (almost surely) has a solution (or equivalenty, the set $\mathcal{W}$ is nonempty) if and only if,*

$$\delta > \delta^* = \delta^*(\kappa) := \inf_{s,r\geq 0} \frac{c_\kappa(s,r)}{r^2}\,. \tag{10}$$

**Remark 2.** *Theorem 1 indicates the necessary and sufficient condition for the existense of GMM. It is worth mentioning that this condition, which is simply the condition on separability of the dataset $\mathcal{D}$, does not depend on the choice of the potential function $\psi(\cdot)$.*

**Remark 3.** *The phase transition (10), is valid for any link function $\rho(\cdot)$. This generalizes the former results in (Candès & Sur, 2018). Note that the summary functional, $c_\kappa(\cdot,\cdot)$, contains the choice of the link function and can be computed numerically.*

The following lemma explains the behavior of $\delta^*$ as $\kappa$ varies.

**Lemma 1.** *$\delta^*$ is a decreasing function of $\kappa$, with $\delta^*(0) = \frac{1}{2}$ and $\lim_{\kappa\to+\infty} \delta^*(\kappa) = 0$.*

The result of Lemma 1 can be intuitively verified. Recall that $\kappa = \frac{\|\mathbf{w}^\star\|}{\sqrt{p}}$ and $y_i \sim \text{RAD}(\rho(\mathbf{x}_i^T\mathbf{w}^\star))$. Therefore, $\kappa \to \infty$ translates to having $y_i = \text{SIGN}(\mathbf{x}_i^T\mathbf{w}^\star)$. In this case our training data is always separable for any number of observations $n$. Besides, the case of $\kappa = 0$ corresponds to having random labels assigned to feature vectors $\mathbf{x}_i$. (Cover, 1965) showed that in this case, as $p \to \infty$, $\delta > 0.5$ is the necessary and sufficient condition for the separability of the dataset.

Figure 1 provides a comparison between the theoretical result in Theorem 1, and the empirical results derived from numerical simulations for $p = 150$ and 20 trials. As seen in this plot, the theory matches well with the empirical simulations.

### 3.2. A nonlinear system of equations

Our main result in Section 3.3 precisely characterizes the performance of GMM in terms of a system of 5 nonlinear equations with 5 unknowns, $(\alpha, \sigma, \beta, \gamma, \tau)$, defined as follows,

$$\begin{cases} \frac{1}{p}\,\mathbb{E}\left[\mathbf{w}^{\star T}\mathbf{P}\right] = \alpha\kappa^2, \\ \frac{1}{p}\,\mathbb{E}\left[\mathbf{h}^T\mathbf{P}\right] = \sqrt{\frac{c_\kappa(\alpha,\sigma)}{\delta}}, \\ \frac{1}{p}\,\mathbb{E}\,\|\mathbf{P}\|^2 = \alpha^2\kappa^2 + \sigma^2, \\ \frac{\partial c_\kappa(\alpha,\sigma)}{\partial\alpha} = \frac{2\kappa^2\gamma}{\beta}\sqrt{c_\kappa(\alpha,\sigma)}, \\ \frac{\partial c_\kappa(\alpha,\sigma)}{\partial\sigma} = \frac{2\sqrt{c_\kappa(\alpha,\sigma)}}{\beta\tau}, \end{cases} \tag{11}$$

where $\mathbf{P}$ is defined as,

$$\mathbf{P} = \text{Prox}_{\sigma\tau\psi(\cdot)}\big((\alpha - \sigma\tau\gamma)\mathbf{w}^\star + \beta\sigma\tau\sqrt{\delta}\mathbf{h}\big) \tag{12}$$

**Remark 4.** *The first three equations in the nonlinear system (11) capture the role of the potential function, via its proximal operator. When $\psi(\cdot)$ is separable, these functions can further be reduced to the proximal operator of a real-valued function. For instance, when $\psi(\cdot) = \|\cdot\|_1$, the proximal operator is simply equivalent to applying the well known shrinkage (defined as $\eta(x,t) = \frac{x}{|x|}(|x| - t)_+$) on each entry. For more information on the proximal operators, please refer to (Parikh et al., 2014).*

### 3.3. Asymptotic performance of GMM

We are now ready to present the main result of the paper. Theorem 2 chracaterizes the asymptotic behavior of GMM, that is the solution to the optimization program (7). It connects the performance of GMM to the solution of the nonlinear system of equations (11), and informally states that,

$$\hat{\mathbf{w}} \xrightarrow{D} \Gamma(\mathbf{w}^\star, \mathbf{h}),\ \text{as } p \to \infty, \tag{13}$$

where $\mathbf{h} \in \mathbb{R}^p$ has standard normal entries, and $\Gamma : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}^p$ is defined as,

$$\Gamma(\mathbf{v}_1, \mathbf{v}_2) = \text{Prox}_{\bar\sigma\bar\tau\psi(\cdot)}\big((\bar\alpha - \bar\sigma\bar\tau\bar\gamma)\mathbf{v}_1 + \bar\beta\bar\sigma\bar\tau\sqrt{\delta}\mathbf{v}_2\big), \tag{14}$$

where $(\bar{\alpha}, \bar{\sigma}, \bar{\beta}, \bar{\gamma}, \bar{\tau})$ is the solution to the nonlinear system (11).

**Theorem 2.** *Let $\hat{\mathbf{w}}$ be the solution of the GMM optimization (7), where for $i = 1, 2, \ldots, n$, $\mathbf{x}_i$ has the multivariate Gaussian distribution $\mathcal{N}(\mathbf{0}, \frac{1}{p}\mathbf{I}_p)$, and $y_i \sim \text{RAD}(\rho(\mathbf{x}_i^T \mathbf{w}^\star))$, and $\mathbf{w}^\star$ is drawn from a distribution $\Pi$ with $\kappa = \frac{\|\mathbf{w}^\star\|}{\sqrt{p}}$. As $n, p \to \infty$ at a fixed overparameterization ratio $\delta = \frac{p}{n} > \delta^*(\kappa)$, the nonlinear system (11) has a unique solution $(\bar{\alpha}, \bar{\sigma}, \bar{\beta}, \bar{\gamma}, \bar{\tau})$. Furthermore, for any locally-Lipschitz function $F : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$, we have,*

$$F(\hat{\mathbf{w}}, \mathbf{w}^\star) \xrightarrow{P} \mathbb{E}[F(\Gamma(\mathbf{w}, \mathbf{h}), \mathbf{w})], \tag{15}$$

*where $\mathbf{h} \in \mathbb{R}^p$ has standard normal entries, $\mathbf{w} \sim \Pi$ is independent of $\mathbf{h}$, and the function $\Gamma(\cdot, \cdot)$ is defined in (14).*

The detailed proof of this result is deferred to Appendix A. In short, we introduce dual variables and write down the Lagrangian which contains a bilinear form with respect to a matrix with i.i.d. Gaussian entries. Exploiting the CGMT framework, we then analyze the nearly-separable auxiliary optimization to find its optimal value, and show that the nonlinear system (11) corresponds to its optimality condition.

**Remark 5.** *The result in Theorem 2 is stated for a general locally-Lipschitz function $F(\cdot, \cdot)$. To evaluate a specific performance measure, one can appeal to this theorem with an appropriate choice of $F$. As an example, the function $F(\mathbf{u}, \mathbf{v}) = \frac{1}{p}\|\mathbf{u} - \mathbf{v}\|^2$ gives the mean-squarred error (MSE).*

### 3.4. Generalization error

Theorem 2 can be utilized to derive useful information on the performance of the classifier. In fact, using this theorem one can show that the parameters $\bar{\alpha}$, and $\bar{\sigma}$ respectively correspond to the correlation (to the underlying parameter) and the mean-squared error of the resulting estimator.

An important measure of performance is the generalization error, which indicates the success of the trained model on unseen data. Here, we compute the generalization error of the GMM classifier. We do so, by appealing to the result of Theorem 2.

**Definition 2.** *The generalization error for a binary classifier with parameter $\hat{\mathbf{w}}$ is defined as,*

$$GE_{\hat{\mathbf{w}}} = \mathbb{P}_{\mathbf{x}}\{\text{SIGN}(\mathbf{x}^T \hat{\mathbf{w}}) \neq \text{SIGN}(\mathbf{x}^T \mathbf{w}^\star)\}, \tag{16}$$

*where the probability is computed with respect to the distribution of the test data.*

It can be shown that when the distribution of the test data is rotationally invariant (e.g., Gaussian, uniform dist. on the unit-sphere), GE only depends on the angle between $\hat{\mathbf{w}}$ and $\mathbf{w}^\star$. The following lemma provides sharp asymptotics on the generalization error of the GMM classifier.

**Lemma 2** (Generalization Error)**.** *Let $\hat{\mathbf{w}}$ be the GMM classifier defined in Section 2.2. Assume $\delta > \delta^*$, and the (test) data is distributed according to the multivariate Gaussian distribution $\mathcal{N}(\mathbf{0}, \frac{1}{p}\mathbf{I}_p)$. Then, as $p \to \infty$, we have,*

$$GE_{\hat{\mathbf{w}}} \xrightarrow{P} \frac{1}{\pi} acos\left(\frac{\kappa\bar{\alpha}}{\sqrt{\kappa^2\bar{\alpha}^2 + \bar{\sigma}^2}}\right), \tag{17}$$

*where $\bar{\alpha}$ and $\bar{\sigma}$ are derived by solving the nonlinear system (11).*

*Proof.* We first note that when the data is normally distributed, the generalization error for $\hat{\mathbf{w}}$ is defined as,

$$GE_{\hat{\mathbf{w}}} = \frac{1}{\pi}\text{acos}\left(\frac{\hat{\mathbf{w}}^T \mathbf{w}^\star}{\|\mathbf{w}^\star\| \, \|\hat{\mathbf{w}}\|}\right). \tag{18}$$

We appeal to the result of Theorem 2 with two different functions. Using $F_1(\mathbf{u}, \mathbf{v}) = \frac{1}{p}\mathbf{v}^T\mathbf{u}$ in (15) will give,

$$\frac{1}{p}\hat{\mathbf{w}}^T \mathbf{w}^\star \xrightarrow{P} \frac{1}{p}\,\mathbb{E}\left[\mathbf{w}^{\star T}\text{Prox}_{\bar{\sigma}\bar{\tau}\psi(\cdot)}\big((\bar{\alpha} - \bar{\sigma}\bar{\tau}\bar{\gamma})\mathbf{w}^\star + \bar{\beta}\bar{\sigma}\bar{\tau}\sqrt{\delta}\mathbf{h}\big)\right]. \tag{19}$$

Since $(\bar{\alpha}, \bar{\sigma}, \bar{\beta}, \bar{\gamma}, \bar{\tau})$ is the solution to the nonlinear system, we can replace the expectation from the first equation in (11),which gives the following,

$$\frac{1}{p}\hat{\mathbf{w}}^T \mathbf{w}^\star \xrightarrow{P} \kappa^2\bar{\alpha}. \tag{20}$$

Similarly, using the result of Theorem 2 for the measure function $F_2(\mathbf{u}, \mathbf{v}) = \frac{1}{p}\|\mathbf{u}\|^2$, along with the third equation in (11) gives,

$$\frac{1}{\sqrt{p}}\|\hat{\mathbf{w}}\| \xrightarrow{P} \sqrt{\kappa^2\bar{\alpha}^2 + \bar{\sigma}^2}\,. \tag{21}$$

The proof is the consequence of (18), (20), and (21), along with the continuity of the function acos$(\cdot)$. □

## 4. GMM for Various Structures

As explained earlier, the potential function $\psi(\cdot)$ is chosen to encourage the structure of the underlying parameter. In this section, we investigate the performance of the GMM classifier for some common structures and the corresponding choices of the potential function.

### 4.1. Max-margin classifier ($\ell_2$-GMM)

The $\ell_2$-norm regularization is commonly used in machine learning applications to stabilize the model. Here, we study the performance of the GMM classifier when $\psi(\cdot) = \frac{1}{2}\|\cdot\|_2^2$, i.e., the solution to the following optimization program,

$$\begin{aligned} \min_{\mathbf{w} \in \mathbb{R}^p} \quad & \frac{1}{2}\|\mathbf{w}\|_2^2 \\ \text{s.t. } & y_i(\mathbf{x}_i^T \mathbf{w}) \geq 1, \quad \text{for } 1 \leq i \leq n. \end{aligned} \tag{22}$$

The optimization program (22) is called the hard-margin SVM and the corresponding solution is the max-margin classifier, as it maximizes the minimum distance (margin) of the datapoints from the separating hyperplane. As mentioned earlier in Section 1, the conventional justification for using such a classifier is that the risk of a classifier is inversely proportional to its margin. The performance of $\ell_2$-GMM (22), has been earlier analyzed in (Deng et al., 2019) and (Montanari et al., 2019). The form we present below in (24), differes in appearance to the results of (Deng et al., 2019), but can be shown to be equivalent.

When $\psi(\cdot)=\frac{1}{2}\|\cdot\|_2^2$, the proximal operator has the following closed-form,

$$\text{Prox}_{\frac{t}{2}\|\cdot\|^2}(\mathbf{u})=\frac{1}{1+t}\mathbf{u}. \tag{23}$$

By replacing the proximal operator in the nonlinear system (11), we can explicitly find two of the variables ($\beta$, and $\gamma$) and reduce it to the following system of three nonlinear equations in three unknowns,

$$\begin{cases} \sqrt{c_\kappa(\alpha,\sigma)}=\sigma\sqrt{\delta}, \\ \dfrac{\partial c_\kappa(\alpha,\sigma)}{\partial\alpha}=\dfrac{-2\kappa^2\alpha\tau\sigma\delta}{1+\sigma\tau}, \\ \dfrac{\partial c_\kappa(\alpha,\sigma)}{\partial\sigma}=\dfrac{2\sigma\delta}{1+\sigma\tau}. \end{cases} \tag{24}$$

### 4.2. Sparse classifier ($\ell_1$-GMM)

In today's machine learning applications, typically the number of available features, $p$, is overwhelmingly large. To reduce the risk of overfitting in such settings, feature selection methods are often performed to exclude irrelevent variables from the model (James et al., 2013). Adding an $\ell_1$ penalty is the most popular approach for feature selection. As a natural consequence of our main result in Theorem 2, here we analyze the asymptotic performance of GMM when the potential function is the $\ell_1$ norm, and evaluate its success on the unseen data (i.e., the test error) when the underlying parameter, $\mathbf{w}^\star$, is sparse.

$$\begin{aligned} \min_{\mathbf{w}\in\mathbb{R}^p} \quad & \|\mathbf{w}\|_1 \\ \text{s.t. } & y_i(\mathbf{x}_i^T\mathbf{w})\geq 1, \quad \text{for } 1\leq i\leq n. \end{aligned} \tag{25}$$

In this case, the proximal operator of the potential function ($\|\cdot\|_1$) is basically equivalent to applying the soft-thresholding operator, on each entry, i.e.,

$$\text{Prox}_{t\|\cdot\|_1}(\mathbf{u})=\eta(\mathbf{u},t), \tag{26}$$

where $\eta(x,t):=\frac{x}{|x|}(|x|-t)_+$ is the soft-thresholding operator. Here, for a sparsity factor $s\in(0,1]$, we assume the entries of $\mathbf{w}^\star$ are sampled i.i.d. from the following distribution,

$$\Pi_s(w)=(1-s)\cdot\delta_0(w)+s\cdot\Big(\frac{\phi(\frac{w}{\frac{\kappa}{\sqrt{s}}})}{\frac{\kappa}{\sqrt{s}}}\Big), \tag{27}$$

where $\delta_0(\cdot)$ is the Dirac delta function, and $\phi(t):=\frac{e^{-\frac{t^2}{2}}}{\sqrt{2\pi}}$ is the density of the standard normal random variable. This means that each of the entries of $\mathbf{w}^\star$ are zero with probability $1-s$, and the nonzero entries have independent Gaussian distribution with variance $\frac{\kappa^2}{s}$. Having this assumption we can further simplify the first three equations in the nonlinear system (11), and present them in terms of q-functions. To streamline our representation, we introduce the following proxies,

$$t_1=\frac{\sigma\tau}{\sqrt{\frac{\kappa^2}{s}(\alpha-\sigma\tau\gamma)^2+\beta^2\sigma^2\tau^2\delta}}, \quad t_2=\frac{1}{\beta\sqrt{\delta}}. \tag{28}$$

We also define the function $\chi:\mathbb{R}\to\mathbb{R}_+$ as,

$$\begin{aligned} \chi(t)&=\mathbb{E}\left[(Z-t)_+^2\right], \quad Z\sim\mathcal{N}(0,1) \\ &=Q(t)(1+t^2)-t\phi(t), \end{aligned} \tag{29}$$

Where $Q(t):=\int_t^\infty\phi(x)dx$ denotes the tail distribution of standard normal random variable. We are now able to simplify the first three equations in (11) and derive the following nonlinear system,

$$\begin{cases} Q(t_1)=\frac{\alpha}{2(\alpha-\sigma\tau\gamma)}, \\ s\cdot Q(t_1)+(1-s)\cdot Q(t_2)=\frac{\sqrt{c_\kappa(\alpha,\sigma)}}{2\beta\sigma\tau\delta}, \\ \frac{s}{t_1^2}\cdot\chi(t_1)+\frac{(1-s)}{t_2^2}\cdot\chi(t_2)=\frac{\kappa^2\alpha^2}{2\sigma^2\tau^2}+\frac{1}{2\tau^2}, \\ \frac{\partial c_\kappa(\alpha,\sigma)}{\partial\alpha}=\frac{2\kappa^2\gamma}{\beta}\sqrt{c_\kappa(\alpha,\sigma)}, \\ \frac{\partial c_\kappa(\alpha,\sigma)}{\partial\sigma}=\frac{2\sqrt{c_\kappa(\alpha,\sigma)}}{\beta\tau}. \end{cases} \tag{30}$$

The nonlinear system (30) can be solved via numerical methods. For our numerical simulations in Section 5 we exploit accelerated fixed-point methods to solve the nonlinear system. Using the the result of Lemma 2, we can compute the generalization error.

Another important measure in this setting (when $\mathbf{w}^\star$ is sparse) is the probability of error in support recovery. Let $\Omega\subseteq[p]$ denote the support of $\mathbf{w}^\star$ (i.e. $\Omega=\{j:\mathbf{w}_j^\star\neq 0\}$.) For a pre-defined threshold $\epsilon$, we form the following estimate of the support,

$$\hat{\Omega}_\epsilon=\{j:1\leq j\leq p,|\hat{\mathbf{w}}_j|>\epsilon\}. \tag{31}$$

The following lemma establishes the success in the support recovery:

**Lemma 3** (Support Recovery)**.** *For a sparsity factor $s\in(0,1]$, let the entries of $\mathbf{w}^\star$ have distribution $\Pi_s$ defined in (27), and $\hat{\mathbf{w}}$ be the solution to the optimization (25). Then, as $p\to\infty$, we have,*

$$\begin{aligned} &\lim_{\epsilon\downarrow 0}P_1(\epsilon):=\mathbb{P}\left\{j\notin\hat{\Omega}_\epsilon|j\in\Omega\right\}\xrightarrow{P}1-2Q(\bar{t}_1) \\ &\lim_{\epsilon\downarrow 0}P_2(\epsilon):=\mathbb{P}\left\{j\in\hat{\Omega}_\epsilon|j\notin\Omega\right\}\xrightarrow{P}2Q(\bar{t}_2)\,, \end{aligned} \tag{32}$$

*where $\bar{t}_1$ and $\bar{t}_2$ are defined as in (28), with variables derived from solving the nonlinear system (30).*

### 4.3. Binary classifier ($\ell_\infty$-GMM)

As the last example of structured classifiers, here we study the case where $\mathbf{w}^\star \in \{\pm\}^p$. To encourage this structure, the potential function is chosen to be the $\ell_\infty$ norm. In linear regression, $\|\cdot\|_\infty$ is used to recover the binary signals, i.e., when $\mathbf{w}^\star \in \{\pm 1\}^p$ (Chandrasekaran et al., 2012). This problem arises in integer programming and has some connections to the Knapsack problem (Mangasarian & Recht, 2011). Here, we consider analyzing the performance of the solution of the following optimization program,

$$\begin{aligned} \min_{\mathbf{w}\in\mathbb{R}^p} \quad & \|\mathbf{w}\|_\infty \\ \text{s.t. } & y_i(\mathbf{x}_i^T\mathbf{w}) \geq 1, \quad \text{for } 1 \leq i \leq n. \end{aligned} \tag{33}$$

It can be shown that the proximal operator of the $\ell_\infty$-norm can be derived by projecting the points onto the $\ell_1$-ball. We use this connection to present the proximal operator in this case in terms of the soft-thresholding operator $\eta(\cdot,\cdot)$.
For a vector $\mathbf{w}$ whose entries are drawn independently from a distribution $\Pi$, we can present the following formula for the proximal operator:

$$\text{Prox}_{tp\|\cdot\|_\infty}(\mathbf{w}) = \mathbf{w} - \text{Prox}_{\lambda\|\cdot\|_1}(\mathbf{w}), \tag{34}$$

where $\lambda := \lambda(t)$ is the smallest nonnegative number that satisfies,

$$\mathbb{E}\left[|\eta(W,\lambda)|\right] = \mathbb{E}\left[(|W|-\lambda)_+\right] \leq t. \tag{35}$$

Here, the expectation is with respect to $W \sim \Pi$. Note that $\lambda$ is a non-increasing function of $t$, and $\lambda = 0$ whenever $t \geq \mathbb{E}\,|W|$.
Similar to the case of $\ell_1$-GMM, here we can use the closed-form of the proximal operator to simplify the first three equations in the nonlinear system (11). For our numerical simulations in the next section, we have done the computations for three different distributions: (1) The i.i.d. Gaussian distribution, (2) the sparse distribution defined in (27), and (3) the uniform binary distribution, $\Pi = \text{Unif}\big(\{\pm 1\}^p\big)$. We postpone the details of the theoretical derivations for this part to Appendix D.3.

## 5. Numerical Simulations

In this section, we investigate the validity of our theoretical results with multiple numerical simulations applied to the three different cases of GMM classifiers elaborated in Section 5. For each of the three potentials discussed in the paper (i.e., $\ell_1$, $\ell_2$, and $\ell_\infty$ norms) we perform numerical simulations for three different models on the distribution of $\mathbf{w}^\star$. In other words, we change the distribution of the entries of $\mathbf{w}^\star$ and evaluate the performance of the aforementioned classifiers on each model. As will observed in our numerical simulations, the appropriate choice of the potential function in the GMM optimization (7) has an impact on the generl-ization error of the resulting classifier. The three different

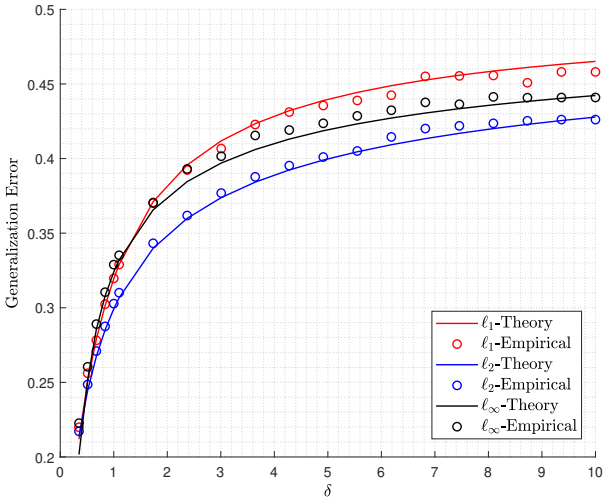


Figure 2: Generalization error of the general max margin classifier under three penalty functions, $\ell_1$ norm with the red line ($\ell_1$-GMM), $\ell_2$ norm with the blue line ($\ell_2$-GMM), and $\ell_\infty$ norm with the black line ($\ell_\infty$-GMM). **In this figure, the entries of $\mathbf{w}^\star$ are drawn independently from $\mathcal{N}(0,\kappa^2)$ Gaussian distribution**. Solid lines correspond to the theoretical results derived from Theorem 2, while the circles are the result of empirical simulations. For the numerical simulations, the result is the average over 100 independent trials with $p = 200$ and $\kappa = 2$.

distribution that we choose for the underlying parameter are as follows:

**Gaussian**: in the first model, we assume that the entries of $\mathbf{w}^\star$ are drawn from a zero-mean Gaussian distribution, $\mathcal{N}(0,\kappa^2)$. In this model, the direction of $\mathbf{w}^\star$ (which indicates the separating hyperplane) is distributed uniformly on the unit sphere. Figure 2 gives the generalization error when $\mathbf{w}^\star$ has Gaussian distribution. The solid lines show the theoretical results derived from Theorem 2 and Lemma 2. The circles depict empirical results that are computed by taking the average over 100 trials with $p = 200$ and $\kappa = 2$. Although our theory provides the generalization error in the asymptotic regime, it appropriately matches the result of empirical simulations in our simulations in finite dimensions. It can be observed in this figure that the max-margin classifier ($\ell_2$-GMM) outperforms the other two classifiers. We should also note that as the overparameterization ratio, $\delta$, grows the generalization error increases which indicates that the estimator is not reliable for large values of $\delta$.
**Sparse**: here, we assume that the entries of $\mathbf{w}^\star$ are drawn from the sparse distribution represented in (27), i.e., each entry is nonzero with probability $s$, and the nonzero entries have i.i.d. Gaussian distribution with appropriately-defined

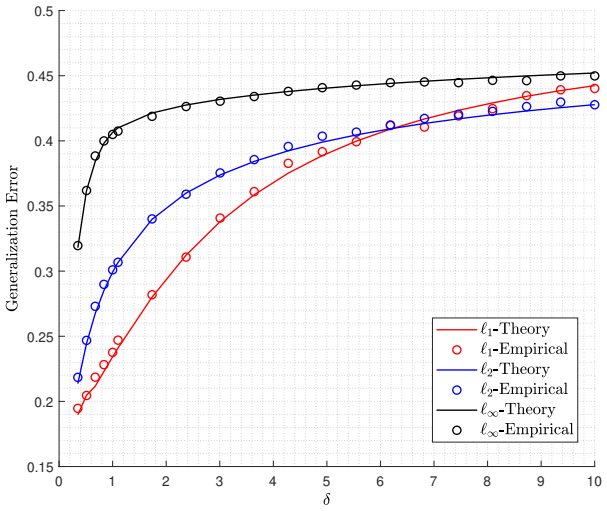


Figure 3: Generalization error of the general max margin classifier under three penalty functions, $\ell_1$ norm with the red line ($\ell_1$-GMM), $\ell_2$ norm with the blue line ($\ell_2$-GMM), and $\ell_\infty$ norm with the black line ($\ell_\infty$-GMM). **In this figure, the underlying vector $\mathbf{w}^\star$ is $s$-sparse, where the non-zero entries are drawn independently from $\mathcal{N}(0, \kappa^2/s)$ Gaussian distribution**. Solid lines correspond to the theoretical results derived from Theorem 2, and the circles are the result of empirical simulations. For the numerical simulations, the result is computed by taking the average over 100 independent trials with $p = 200$, $s = .1$ and $\kappa = 2$.

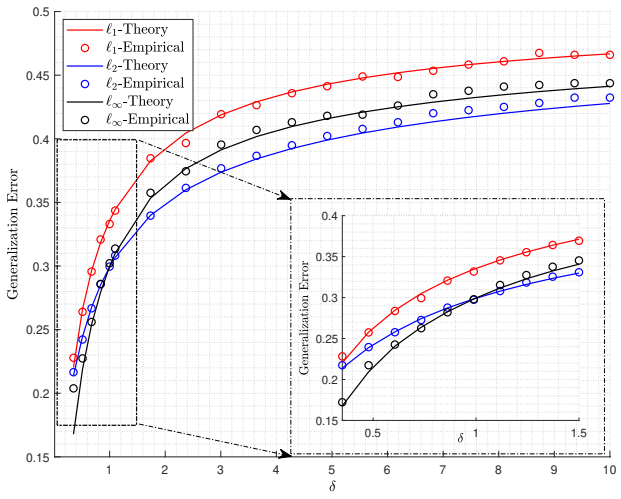


Figure 4: Generalization error of the general max margin classifier under three penalty functions, $\ell_1$ norm with the red line ($\ell_1$-GMM), $\ell_2$ norm with the blue line ($\ell_2$-GMM), and $\ell_\infty$ norm with the black line ($\ell_\infty$-GMM). **In this figure, the entries of $\mathbf{w}^\star$ are drawn independently from $\kappa * \textsc{Rad}(0.5)$ Rademacher distribution**. Solid lines correspond to the theoretical results derived from Theorem 2, and the circles are the result of empirical simulations. For the numerical simulations, the result is the average over 100 independend trials with $p = 200$ and $\kappa = 2$.

variance. Figure 3 demonstrates the result of the numerical simulations for this model for the three different classifiers of interest. The empirical result is the average over 100 trials with $p = 200$, $s = 0.1$, and $\kappa = 2$. Similar to the previous case, the empirical results match the theory. Also, it can be observed that the $\ell_1$-GMM outperforms the two other classifiers in the regime of $\delta$ that the classifiers performs well (i.e. $\delta \lessapprox 6$.) Similarly, we can observe that for large values of $\delta$ all the classifiers perform poorly.

**Binary**: in this model the entries of $\mathbf{w}^\star$ are independently drawn from $\{+\kappa, -\kappa\}$, i.e., $\mathbf{w}^\star$ is uniformly chosen on the discrete set $\{\pm\kappa\}^p$. Figure 4 shows the result of numerical simulations under this model. Similar to previous cases the empirical results ($\kappa = 2,\ p = 200$) match the theory. Also, the $\ell_\infty$-GMM classifier outperforms the other two classifiers for $\delta < 1$ (which corresponds to the underparameterized setting). However, the max-margin classifier performs better for larger values of $\delta$.

## 6. Conclusion and Future Directions

In this paper, we introduced the generalized margin maximizers (GMM) as a way to extend the max-margin classifiers to structured models. To this end, we proposed an optimization program whose objective function is a convex potential function $\psi(\cdot)$ that encourages the underlying structure, and the constraints are similar to the max-margin classifier (hard-margin SVM). Our main result in Theorem 2 provides the asymptotic behavior of GMM classifier for any locally-Lipschitz performance measure via solving a system of nonlinear equations. We utilize this result to characterize the generalization error in the asymptotic regime.

We examined our theoretical findings on three specific choices of the potential function, $\ell_1$, $\ell_2$, and $\ell_\infty$ norms. We simplified the nonlinear systems for each of these functions and validated our theoretical results in numerical simulations by doing simulations on three different structures on the underlying parameter, $\mathbf{w}^*$. The numerical simulations indicates that for sparse signals, $\ell_1$-GMM outperforms the max-margin classifier ($\ell_2$-GMM). We also observed that for binary signals, when $\delta < 1$, the $\ell_\infty$-GMM outperforms the two other classifiers.

In future works, we would like to extend our theory to predict some common phenomena (e.g. the double descent) for GMM. Also, another avenue of pursuit is to design iterative optimizaiton algorithms that would converge to the GMM classifier.

## References


Abbasi, E., Salehi, F., and Hassibi, B. Performance analysis of convex data detection in mimo. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4554–4558. IEEE, 2019.

Atitallah, I. B., Thrampoulidis, C., Kammoun, A., Al-Naffouri, T. Y., Hassibi, B., and Alouini, M.-S. Ber analysis of regularized least squares for bpsk recovery. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4262–4266. IEEE, 2017.

Bartlett, P. L., Long, P. M., Lugosi, G., and Tsigler, A. Benign overfitting in linear regression. *arXiv preprint arXiv:1906.11300*, 2019.

Belkin, M., Hsu, D. J., and Mitra, P. Overfitting or perfect fitting? risk bounds for classification and regression rules that interpolate. In *Advances in neural information processing systems*, pp. 2300–2311, 2018.

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

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

Chandrasekaran, V., Recht, B., Parrilo, P. A., and Willsky, A. S. The convex geometry of linear inverse problems. *Foundations of Computational mathematics*, 12(6):805–849, 2012.

Cortes, C. and Vapnik, V. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

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

Deng, Z., Kammoun, A., and Thrampoulidis, C. A model of double descent for high-dimensional binary linear classification. *arXiv preprint arXiv:1911.05822*, 2019.

Dhifallah, O., Thrampoulidis, C., and Lu, Y. M. Phase retrieval via polytope optimization: Geometry, phase transitions, and new algorithms. *arXiv preprint arXiv:1805.09555*, 2018.

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

Hastie, T., Montanari, A., Rosset, S., and Tibshirani, R. J. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*, 2019.

James, G., Witten, D., Hastie, T., and Tibshirani, R. *An introduction to statistical learning*, volume 112. Springer, 2013.

Jourani, A., Thibault, L., and Zagrodny, D. Differential properties of the moreau envelope. *Journal of Functional Analysis*, 266(3):1185–1237, 2014.

Kini, G. and Thrampoulidis, C. Analytic study of double descent in binary classification: The impact of loss. *arXiv preprint arXiv:2001.11572*, 2020.

Mangasarian, O. L. and Recht, B. Probability of unique integer solution to a system of linear equations. *European Journal of Operational Research*, 214(1):27–30, 2011.

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

Miolane, L. and Montanari, A. The distribution of the lasso: Uniform control over sparse balls and adaptive parameter tuning. *arXiv preprint arXiv:1811.01212*, 2018.

Montanari, A., Ruan, F., Sohn, Y., and Yan, J. The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime. *arXiv preprint arXiv:1911.01544*, 2019.

Parikh, N., Boyd, S., et al. Proximal algorithms. *Foundations and Trends® in Optimization*, 1(3):127–239, 2014.

Salehi, F., Abbasi, E., and Hassibi, B. Learning without the phase: Regularized phasemax achieves optimal sample complexity. In *Advances in Neural Information Processing Systems*, pp. 8641–8652, 2018a.

Salehi, F., Abbasi, E., and Hassibi, B. A precise analysis of phasemax in phase retrieval. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pp. 976–980. IEEE, 2018b.

Salehi, F., Abbasi, E., and Hassibi, B. The impact of regularization on high-dimensional logistic regression. In *Advances in Neural Information Processing Systems*, pp. 11982–11992, 2019.

Soudry, D., Hoffer, E., Nacson, M. S., Gunasekar, S., and Srebro, N. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1):2822–2878, 2018.

Stojnic, M. A framework to characterize performance of lasso algorithms. *arXiv preprint arXiv:1303.7291*, 2013.

Sur, P. and Candès, E. J. A modern maximum-likelihood theory for high-dimensional logistic regression. *arXiv preprint arXiv:1803.06964*, 2018.

Taheri, H., Pedarsani, R., and Thrampoulidis, C. Sharp guarantees for solving random equations with one-bit information. In *2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 765–772. IEEE, 2019.

Thrampoulidis, C. *Recovering structured signals in high dimensions via non-smooth convex optimization: Precise performance analysis*. PhD thesis, California Institute of Technology, 2016.

Thrampoulidis, C., Oymak, S., and Hassibi, B. Regularized linear regression: A precise analysis of the estimation error. In *Conference on Learning Theory*, pp. 1683–1709, 2015.

Thrampoulidis, C., Abbasi, E., and Hassibi, B. Precise error analysis of regularized $m$-estimators in high dimensions. *IEEE Transactions on Information Theory*, 64(8):5592–5628, 2018.

Thrampoulidis, C., Zadik, I., and Polyanskiy, Y. A simple bound on the ber of the map decoder for massive mimo systems. *arXiv preprint arXiv:1903.03949*, 2019.

Vapnik, V. Estimation of dependences based on empirical data berlin, 1982.

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

Xu, J. and Hsu, D. How many variables should be entered in a principal component regression equation? *arXiv preprint arXiv:1906.01139*, 2019.

# Appendix

## A. Proof of Theorem 2

Here, we present the proof of the main result of the paper. Recall that the generalized margin maximizer is defined as the soluiton to the following optimization program,

$$\begin{aligned} \min_{\mathbf{w}\in\mathbb{R}^p} \quad & \psi(\mathbf{w}) \\ \text{s.t. } & y_i(\mathbf{x}_i^T\mathbf{w}) \geq 1, \quad \text{for } 1 \leq i \leq n. \end{aligned} \tag{36}$$

Theorem 2 provides a precise characterization on the performance of this optimization program in the asymptotic regime, where $n, p \to \infty$ at a fixed ratio $\delta := n/p$. We assume the datapoints are drawn independently from the multivariate gaussian distribution, i.e., $\mathbf{x}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \frac{1}{p}\mathbf{I}_p)$.

For our analysis we utilize the CGMT framework (see Appendix E.2), which will provide us with a nearly-separable optimization program that has the same performance as (36). To simplify the presentation, we are breaking down the proof into the following three steps:

1. Finding the auxiliary optimization: By introducing dual variables, we present the optimization (36) as a bilinear form with respect to a Gaussian matrix. Consequently, we use the result of Theorem 3 and Corollary 1 to find the auxiliary optimization.
2. Analyzing the auxiliary optimization: The first step provides a nearly-separable optimization. The purpose of this step is to simplify this optimization and present it in terms of an optimization program with respect to scalar variables.
3. Optimality condition of the auxiliary optimization: By taking the derivatives with respect to various scalars, we present the first-order optimality condition on the solution of the (simplified) auxiliary optimization. Further simplification gives the nonlinear system (11).

We explain each of the three steps in more details in the following subsections.

### A.1. Finding the auxiliary optimization

The following lemma presents the auxiliary optimization associated with the GMM optimization (36).

**Lemma 4.** *Let $\hat{\mathbf{w}}$ be the solution to the optimization (36). Consider the following optimization:*

$$\begin{aligned} \min_{\substack{\alpha\in\mathbb{R} \\ \tilde{\mathbf{w}}\in\mathbb{R}^p \\ \tilde{\mathbf{w}}\perp\mathbf{w}^\star}} \quad & \frac{1}{p}\psi(\alpha\mathbf{w}^\star + \tilde{\mathbf{w}}) \\ \textit{s.t.} \quad & \frac{1}{p}(\mathbf{h}^T\tilde{\mathbf{w}})^2 \geq n \cdot c_\kappa\big(\alpha, \frac{\|\tilde{\mathbf{w}}\|}{\sqrt{p}}\big), \end{aligned} \tag{37}$$

*where $\mathbf{h} \in \mathbb{R}^p$ has i.i.d. standard normal entries. Assume $(\bar{\alpha}, \bar{\tilde{\mathbf{w}}}) \in \mathbb{R} \times \mathbb{R}^p$ be the solution to this optimization program. Then, as $p \to \infty$, we have:*

$$\big\|\hat{\mathbf{w}} - (\bar{\alpha}\mathbf{w}^\star + \bar{\tilde{\mathbf{w}}})\big\| \xrightarrow{P} 0. \tag{38}$$

*Proof.* In order to apply the CGMT, we need to have a min-max optimization. Introducing the Lagrange variable, $\boldsymbol{\lambda} := [\lambda_1, \lambda_2, \ldots, \lambda_n]^T \in \mathbb{R}^n_+$, we can rewrite the optimization program as follows,

$$\min_{\mathbf{w}\in\mathbb{R}^p} \max_{\boldsymbol{\lambda}\in\mathbb{R}^n_+} \frac{1}{p}\psi(\mathbf{w}) + \frac{1}{n}\sum_{i=1}^n \lambda_i\big(1 - y_i(\mathbf{x}_i^T\mathbf{w})\big). \tag{39}$$

Note that the scaling has been performed in such a way that all the terms in the objective be of constant order. We define the matrix $\mathbf{H} \in \mathbb{R}^{n\times p}$ as,

$$\mathbf{H} := -\sqrt{p} \cdot \begin{bmatrix} -\mathbf{x}_1^T- \\ -\mathbf{x}_2^T- \\ \vdots \\ -\mathbf{x}_n^T- \end{bmatrix}. \tag{40}$$

Based on the assumption on the distribution of datapoints, this matrix has i.i.d. standrad normal $\mathcal{N}(0,1)$ entries. To ease the notation, we also define a new variable $\bar{\boldsymbol{\lambda}} = \boldsymbol{\lambda} \odot \mathbf{y}$ (i.e., $\bar{\lambda}_i = \lambda_i y_i$) and reformulate the optimization (39) as,

$$\min_{\mathbf{w}\in\mathbb{R}^p} \max_{\substack{\bar{\boldsymbol{\lambda}}\in\mathbb{R}^n \\ \bar{\lambda}_i y_i \geq 0}} \frac{1}{p}\psi(\mathbf{w}) + \frac{1}{n}\bar{\boldsymbol{\lambda}}^T\mathbf{y} + \frac{1}{n\sqrt{p}}\bar{\boldsymbol{\lambda}}^T\mathbf{H}\mathbf{w}. \tag{41}$$

We proceed by analyzing the optimization program (41). In order to apply the CGMT, we need an additive bilinear form that is statistically independent of other functions that appear in the objective. Note that the label vector $\mathbf{y} \in \{\pm\}^n$ is a random variables that depends on $\mathbf{H}\mathbf{w}^\star$, as $\mathbf{y} = \textsc{Rad}\big(\rho(-\frac{1}{\sqrt{p}}\mathbf{H}\mathbf{w}^\star)\big)$. Therefore, to remove this independence between $\mathbf{y}$ and the bilinear form, we use the projection onto $\mathbf{w}^\star$. Let $\mathbf{P}$ be the matrix of orthogonal projection onto span$(\mathbf{w}^\star)$, i.e., $\mathbf{P} = \frac{1}{\|\mathbf{w}^\star\|^2}\mathbf{w}^\star\mathbf{w}^{\star T}$, and $\mathbf{P}^\perp$ be its orthogonal completent, $\mathbf{P}^\perp = \mathbf{I}_p - \mathbf{P}$. We use these projection matrices to decompose the Gaussian matrix $\mathbf{H}$ as $\mathbf{H} = \mathbf{H}_1 + \mathbf{H}_2$ with $\mathbf{H}_1 := \mathbf{H}\times\mathbf{P}$, and $\mathbf{H}_2 := \mathbf{H}\times\mathbf{P}^\perp$. This gives the following equivalent optimization,

$$\min_{\mathbf{w}\in\mathbb{R}^p} \max_{\substack{\bar{\boldsymbol{\lambda}}\in\mathbb{R}^n \\ \bar{\lambda}_i y_i \geq 0}} \frac{1}{p}\psi(\mathbf{w}) + \frac{1}{n}\bar{\boldsymbol{\lambda}}^T\mathbf{y} + \frac{1}{n\sqrt{p}}\bar{\boldsymbol{\lambda}}^T\mathbf{H}_1\mathbf{w} + \frac{1}{n\sqrt{p}}\bar{\boldsymbol{\lambda}}^T\mathbf{H}_2\mathbf{w}. \tag{PO}$$

It is worth noting that the projections of a Gaussian matrix (or vector) onto orthogonal subspaces are statistically independent. Also, the label vector $\mathbf{y}$ would be independent of $\mathbf{H}_2$ since,

$$\mathbf{y} = \textsc{Rad}\big(\rho(-\frac{1}{\sqrt{p}}\mathbf{H}\mathbf{w}^\star)\big) = \textsc{Rad}\big(\rho(-\frac{1}{\sqrt{p}}\mathbf{H}\mathbf{P}\mathbf{w}^\star)\big) = \textsc{Rad}\big(\rho(-\frac{1}{\sqrt{p}}\mathbf{H}_1\mathbf{w}^\star)\big), \tag{42}$$

where we used $\mathbf{P}\mathbf{w}^\star = \mathbf{w}^\star$. Therefore, all the additive terms in the objective function of (PO) except the last one are independent of $\mathbf{H}_2$. Also, the objective function is convex with respect to $\mathbf{w}$ and concave(linear) with respect to $\bar{\boldsymbol{\lambda}}$. In order to apply the CGMT framework (Theorem 3), we only need an extra condition which is restricting the feasibility sets of $\mathbf{w}$, and $\bar{\boldsymbol{\lambda}}$ to be compact and convex. We can introduce some artificial convex and bounded sets $\mathcal{S}_{\mathbf{w}}$, and $\mathcal{S}_{\bar{\boldsymbol{\lambda}}}$, and perform the optimization over these sets. Note that these sets can be chosen large enough such that they do not affect the optimization itself. For simplicity, in our arguments here we ignore the condition on the compactness of the fesible sets and apply the CGMT whenever the variables are defined on a convex domain.

The optimization program (PO) is suitable to be analyzed via the CGMT as the conditions are all satisfied. Having identified (PO) as the primary optimization, it is straightforward to write its corresponding auxiliary optimization (AO) [as in (96), c.f. Appendix E.2]. The Auxiliary Optimization (AO) can be written as follows,

$$\min_{\mathbf{w}\in\mathbb{R}^p} \max_{\substack{\bar{\boldsymbol{\lambda}}\in\mathbb{R}^n \\ \bar{\lambda}_i y_i \geq 0}} \frac{1}{p}\psi(\mathbf{w}) + \frac{1}{n}\bar{\boldsymbol{\lambda}}^T\mathbf{y} + \frac{1}{n\sqrt{p}}\bar{\boldsymbol{\lambda}}^T\mathbf{H}_1\mathbf{w} + \frac{1}{n\sqrt{p}}\big(\,\|\bar{\boldsymbol{\lambda}}\|\,\mathbf{h}^T\mathbf{P}^\perp\mathbf{w} + \bar{\boldsymbol{\lambda}}^T\mathbf{g}\,\|\mathbf{P}^\perp\mathbf{w}\|\,\big), \tag{AO}$$

where $\mathbf{h}\in\mathbb{R}^p$ and $\mathbf{g}\in\mathbb{R}^n$ have i.i.d. standard normal entries. Next, we decompose $\mathbf{w}$ as $\mathbf{w} := \mathbf{P}\mathbf{w} + \mathbf{P}^\perp\mathbf{w} = \alpha\mathbf{w}^\star + \tilde{\mathbf{w}}$, where $\alpha\in\mathbb{R}$, and $\tilde{\mathbf{w}}\in\mathbb{R}^p$ is such that $\tilde{\mathbf{w}}\perp\mathbf{w}^\star$. We also define the vector $\mathbf{q} := -\frac{1}{\kappa\sqrt{p}}\mathbf{H}\mathbf{w}^\star$. Note that since $\|\mathbf{w}\| = \kappa\sqrt{p}$, the entries of $\mathbf{q}$ have standrad normal distribution. Therefore, we have the following equivalent optimization,

$$\min_{\substack{\alpha\in\mathbb{R} \\ \tilde{\mathbf{w}}\in\mathbb{R}^p \\ \tilde{\mathbf{w}}\perp\mathbf{w}^\star}} \max_{\substack{\bar{\boldsymbol{\lambda}}\in\mathbb{R}^n \\ \bar{\lambda}_i y_i \geq 0}} \frac{1}{p}\psi(\alpha\mathbf{w}^\star + \tilde{\mathbf{w}}) + \frac{1}{n}\bar{\boldsymbol{\lambda}}^T\mathbf{y} - \frac{\alpha\kappa}{n}\bar{\boldsymbol{\lambda}}^T\mathbf{q} + \frac{1}{n\sqrt{p}}\big(\,\|\bar{\boldsymbol{\lambda}}\|\,\mathbf{h}^T\tilde{\mathbf{w}} + \bar{\boldsymbol{\lambda}}^T\mathbf{g}\,\|\tilde{\mathbf{w}}\|\,\big)\,, \tag{43}$$

Proceeding onwards, we solve the inner optimization ($\max_{\bar{\boldsymbol{\lambda}}}$) with respect to the direction of $\bar{\boldsymbol{\lambda}}$. We have:

$$\max_{\substack{\bar{\boldsymbol{\lambda}}\in\mathbb{R}^n \\ \bar{\lambda}_i y_i \geq 0}} \frac{1}{n}\bar{\boldsymbol{\lambda}}^T\mathbf{y} - \frac{\alpha\kappa}{n}\bar{\boldsymbol{\lambda}}^T\mathbf{q} + \frac{1}{n\sqrt{p}}\big(\,\|\bar{\boldsymbol{\lambda}}\|\,\mathbf{h}^T\tilde{\mathbf{w}} + \bar{\boldsymbol{\lambda}}^T\mathbf{g}\,\|\tilde{\mathbf{w}}\|\,\big) = \max_{\substack{\bar{\boldsymbol{\lambda}}\in\mathbb{R}^n \\ \bar{\lambda}_i y_i \geq 0}} \frac{\|\bar{\boldsymbol{\lambda}}\|}{\sqrt{n}}\big(\frac{1}{\sqrt{np}}\mathbf{h}^T\tilde{\mathbf{w}} + \frac{1}{\sqrt{n}}\,\|\boldsymbol{\mu}\|\big) \tag{44}$$

$$\text{s.t. } \mu_i = \Big(1 - \alpha\kappa q_i y_i + \frac{\|\tilde{\mathbf{w}}\|}{\sqrt{p}}g_i y_i\Big)_+, \text{ for } 1\leq i\leq n.$$

Recall that the function $c_\kappa : \mathbb{R}\times\mathbb{R}_+ \to \mathbb{R}$ is defined (c.f. Definition 1) as follows:

$$c_\kappa(t_1,t_2) = \mathbb{E}\,\big(1 - \kappa t_1 Z_1 Y + t_2 Z_2\big)_+^2, \tag{45}$$

where $Z_1, Z_2$ are independent standard normal random variables, and $Y \sim \text{RAD}\big(\rho(\kappa Z_1)\big)$. Therefore, we have $\mathbb{E}\,\mu_i^2 = c_\kappa(\alpha, \frac{\|\tilde{\mathbf{w}}\|}{\sqrt{p}})$, and using the SLLN, as $p, n \to \infty$ we can replace $\|\boldsymbol{\mu}\|$ with $\sqrt{n \cdot c_\kappa(\alpha, \frac{\|\tilde{\mathbf{w}}\|}{\sqrt{p}})}$ due to the almost sure convergence. Introducing the positive variable $\beta = \frac{\|\bar{\boldsymbol{\lambda}}\|}{\sqrt{n}}$, we have the following reformulation of the auxiliary optimization,

$$\min_{\substack{\alpha\in\mathbb{R}\\ \tilde{\mathbf{w}}\in\mathbb{R}^p\\ \tilde{\mathbf{w}}\perp\mathbf{w}^\star}} \max_{\beta\geq 0} \frac{1}{p}\psi(\alpha\mathbf{w}^\star + \tilde{\mathbf{w}}) + \frac{\beta}{\sqrt{np}}\mathbf{h}^T\tilde{\mathbf{w}} + \beta\cdot\sqrt{c_\kappa\big(\alpha, \frac{\|\tilde{\mathbf{w}}\|}{\sqrt{p}}\big)}\ . \tag{46}$$

We can write the inner maximization (with respect to $\beta$) as a constraint for the optimization, which gives the same formulation as (37), i.e.,

$$\begin{aligned}\min_{\substack{\alpha\in\mathbb{R}\\ \tilde{\mathbf{w}}\in\mathbb{R}^p\\ \tilde{\mathbf{w}}\perp\mathbf{w}^\star}} \ & \frac{1}{p}\psi(\alpha\mathbf{w}^\star + \tilde{\mathbf{w}}) \\ \text{s.t.} \ & \frac{1}{p}(\mathbf{h}^T\tilde{\mathbf{w}})^2 \geq n\cdot c_\kappa\big(\alpha, \frac{\|\tilde{\mathbf{w}}\|}{\sqrt{p}}\big),\end{aligned} \tag{47}$$

Using the result of Corollary 1, we have that when the solution of the primary optimization converges as the problem dimensions grow ($p \to \infty$), the solution of the auxiliary optimization converges to the same set (point). This concludes the proof. □

### A.2. Analyzing the auxiliary optimization

In this section we analyze the performance of the refined version of the auxiliary optimization in (46). Although this optimization program is (nearly) separable, it is still a high-dimensional optimization. Ideally, one would like to simplify this optimzation to obtain another optimization program in lower dimensions (with respect to a few scalar variables) where the performance can be numerically computed. To do so, in this section we exploit some tools from convex analysis along with some tricks from calculus to further simplify the optimization program (46).

The goal is to express the final result in terms of the *expected Moreau envelope* of the regularization function. To better understand the behavior of the solution in (46) we first introduce some new variables, $\mathbf{u}, \mathbf{v} \in \mathbb{R}^p$, and $\gamma \in \mathbb{R}$ and write the optimization as follows,

$$\min_{\substack{\alpha\in\mathbb{R}\\ \mathbf{u},\tilde{\mathbf{w}}\in\mathbb{R}^p}} \max_{\substack{\beta\geq 0,\gamma\\ \mathbf{v}\in\mathbb{R}^p}} \frac{1}{p}\psi(\mathbf{u}) + \frac{\beta}{\sqrt{np}}\mathbf{h}^T\tilde{\mathbf{w}} + \beta\cdot\sqrt{c_\kappa\big(\alpha, \frac{\|\tilde{\mathbf{w}}\|}{\sqrt{p}}\big)} + \frac{1}{p}\mathbf{v}^T\big(\mathbf{u} - \alpha\mathbf{w}^\star - \tilde{\mathbf{w}}\big) + \frac{\gamma}{p}\mathbf{w}^{\star T}\tilde{\mathbf{w}}. \tag{48}$$

The variable $\mathbf{u}$ has been introduced to detach the impact of $s$ and $\tilde{\mathbf{w}}$ from $\psi(\cdot)$. The variables $\mathbf{v}$ and $\gamma$ are Lagrange dual variables to remove the constraints from the optimization. We shall emphasize again that the normalization has been performed to ensure that all the terms in the objective are of constant order. Next, we would like to solve the minimization with respect to $\tilde{\mathbf{w}}$.

Before continuing our analysis, we need to discuss an important point that would help us in the remaining of this section. It will be observed that in order to simplify the optimization, we would like to flip the orders of $\min$ and $\max$ in the (AO) optimization. Since the objective function in the auxiliary optimization is not convex-concave we cannot appeal to the Sion's min-max theorem in order to flip $\min$ and $\max$. However, it has been shown (see Appendix A in (Thrampoulidis et al., 2018)) that flipping the orders of $\min$ and $\max$ in the (AO) is allowed in the asymptotic setting. This is mainly due to the fact that the original (PO) optimization was convex-concave with respect to its variables, and as the CGMT suggests (AO) and (PO) are tightly related in the asymptotic setting; hence, flipping the order of optimizations in (AO) is justified whenever such a flipping is allowed in the (PO). We appeal to this result to flip the orders of $\min$ and $\max$ when needed.

Next, we solve the optimization with respect to the direction of $\tilde{\mathbf{w}}$. Defining $\sigma := \|\tilde{\mathbf{w}}\|/\sqrt{p}$ and solving the optimization with respect to the direction of $\tilde{\mathbf{w}}$ leads to,

$$\min_{\substack{\sigma\geq 0,\alpha\\ \mathbf{u}\in\mathbb{R}^p}} \max_{\substack{\beta\geq 0,\gamma\\ \mathbf{v}\in\mathbb{R}^p}} \frac{1}{p}\psi(\mathbf{u}) + \beta\cdot\sqrt{c_\kappa\big(\alpha,\sigma\big)} + \frac{1}{p}\mathbf{v}^T\big(\mathbf{u} - \alpha\mathbf{w}^\star\big) - \sigma\cdot\left\|\frac{\beta}{\sqrt{n}}\mathbf{h} - \frac{1}{\sqrt{p}}\mathbf{v} + \frac{\gamma}{\sqrt{p}}\mathbf{w}^\star\right\|. \tag{49}$$

Consequently, we are considering the maximization with respect to the vector variable $\mathbf{v} \in \mathbb{R}^p$. As seen in (49) this variable appears in the last two additive terms in the objective function. To find the optimal value for $\mathbf{v}$, we introduce a new scalar variable $\tau > 0$ [3], which simplifies the optimization by changing $\|\cdot\|$ to $\|\cdot\|^2$. The new optimization would be,

$$\min_{\substack{\sigma\geq 0,\alpha\\ \mathbf{u}\in\mathbb{R}^p}} \max_{\substack{\beta\geq 0,\tau>0,\gamma\\ \mathbf{v}\in\mathbb{R}^p}} \frac{1}{p}\psi(\mathbf{u}) + \beta\cdot\sqrt{c_\kappa(\alpha,\sigma)} + \frac{1}{p}\mathbf{v}^T\big(\mathbf{u}-\alpha\mathbf{w}^\star\big) - \frac{\sigma\tau}{2} - \frac{\sigma}{2\tau}\cdot\left\|\frac{\beta}{\sqrt{n}}\mathbf{h} - \frac{1}{\sqrt{p}}\mathbf{v} + \frac{\gamma}{\sqrt{p}}\mathbf{w}^\star\right\|^2. \tag{50}$$

It can be easily check that the optimization programs (49) and (50) are equivalent by simply solving the inner optimization with respect to the variable $\tau$. We are now ready to solve the optimization with respect to $\mathbf{v}$. To do so, we continue by making a completion of squares as follows,

$$\begin{aligned}
\frac{1}{p}\mathbf{v}^T\big(\mathbf{u}-\alpha\mathbf{w}^\star\big) - \frac{\sigma}{2\tau}\cdot\left\|\frac{\beta}{\sqrt{n}}\mathbf{h} - \frac{1}{\sqrt{p}}\mathbf{v} + \frac{\gamma}{\sqrt{p}}\mathbf{w}^\star\right\|^2 &= -\frac{\sigma}{2\tau}\cdot\left\|\frac{\beta}{\sqrt{n}}\mathbf{h} - \frac{1}{\sqrt{p}}\mathbf{v} + \frac{\gamma}{\sqrt{p}}\mathbf{w}^\star + \frac{\tau}{\sigma\sqrt{p}}\mathbf{u} - \frac{\alpha\tau}{\sigma\sqrt{p}}\mathbf{w}^\star\right\|^2 \\
&\quad + \frac{\tau}{2\sigma p}\|\mathbf{u}-\alpha\mathbf{w}^\star\|^2 + \frac{\beta}{\sqrt{np}}\mathbf{u}^T\mathbf{h} + \frac{\gamma}{p}\mathbf{u}^T\mathbf{w}^\star - \frac{\alpha\beta\sqrt{\delta}}{p}\mathbf{h}^T\mathbf{w}^\star - \alpha\gamma\kappa^2, \\
\boxed{p,n\to+\infty} \qquad &= -\frac{\sigma}{2\tau}\cdot\left\|\frac{\beta}{\sqrt{n}}\mathbf{h} - \frac{1}{\sqrt{p}}\mathbf{v} + \frac{\gamma}{\sqrt{p}}\mathbf{w}^\star + \frac{\tau}{\sigma\sqrt{p}}\mathbf{u} - \frac{\alpha\tau}{\sigma\sqrt{p}}\mathbf{w}^\star\right\|^2 \\
&\quad + \frac{\tau}{2\sigma p}\left\|\mathbf{u} + \frac{\beta\sigma\sqrt{\delta}}{\tau}\,\mathbf{h} + (\frac{\sigma\gamma}{\tau}-\alpha)\mathbf{w}^\star\right\|^2 - \frac{\sigma}{2\tau}(\delta\beta^2+\gamma^2\kappa^2),
\end{aligned} \tag{51}$$

Where we exploit the fact that, as $p\to\infty$, we can replace $\frac{1}{p}\|\mathbf{w}^\star\|^2$ and $\frac{1}{p}\|\mathbf{h}\|^2$ with $\kappa^2$ and 1, respectively. Furthermore, we omit the term $\frac{1}{p}\mathbf{h}^T\mathbf{w}^\star = \mathcal{O}(\frac{1}{\sqrt{p}})$ as its negligible compare to other terms in the optimization (which are of constant $\mathcal{O}(1)$ orders.) Using the above completion-of-squares $\mathbf{v}$ is now appearing in only one quadratic term in (51). Hence, To maximize the objective, $\mathbf{v}$ chooses itself in such a way that it makes the quadratic term equal to zero. This gives the following optimization,

$$\min_{\substack{\sigma\geq 0,\alpha\\ \mathbf{u}\in\mathbb{R}^p}} \max_{\substack{\beta\geq 0,\tau>0,\gamma\\ \mathbf{v}\in\mathbb{R}^p}} \beta\cdot\sqrt{c_\kappa(\alpha,\sigma)} - \frac{\sigma\tau}{2} - \frac{\sigma}{2\tau}\big(\delta\beta^2+\gamma^2\kappa^2\big) + \frac{1}{p}\big[\psi(\mathbf{u}) + \frac{\tau}{2\sigma}\left\|\mathbf{u} + \frac{\beta\sigma\sqrt{\delta}}{\tau}\,\mathbf{h} + (\frac{\sigma\gamma}{\tau}-\alpha)\mathbf{w}^\star\right\|^2\big]. \tag{52}$$

We now switch the order of $\min$ and $\max$ (similar to what we did earlier for $\tilde{\mathbf{w}}$) and perform the minimization with respect to $\mathbf{u}$. Using the definition of the Moreau envelope, we can write down this optimization in terms of the Moreau envelpe of the potential function. We have,

$$M_{\psi(\cdot)}\big((\alpha-\frac{\sigma\gamma}{\tau})\mathbf{w}^\star - \frac{\beta\sigma\sqrt{\delta}}{\tau}\,\mathbf{h}, \frac{\sigma}{\tau}\big) \;=\; \min_{u\in\mathbb{R}^p}\; \psi(\mathbf{u}) + \frac{\tau}{2\sigma}\left\|\mathbf{u} + \frac{\beta\sigma\sqrt{\delta}}{\tau}\,\mathbf{h} + (\frac{\sigma\gamma}{\tau}-\alpha)\mathbf{w}^\star\right\|^2. \tag{53}$$

Using the result of Lemma 9, we have that the Moreau envelope is a Lipschitz function as $\psi(\cdot)$ is Lipschitz. Therefore, we can exploit the Gaussian concentration of Lipschitz functions (see Theorem 5.22 in (Vershynin, 2018)) which gives,

$$\frac{1}{p}M_{\psi(\cdot)}\big((\alpha-\frac{\sigma\gamma}{\tau})\mathbf{w}^\star - \frac{\beta\sigma\sqrt{\delta}}{\tau}\,\mathbf{h}, \frac{\sigma}{\tau}\big) \xrightarrow{P} \frac{1}{p}\,\mathbb{E}\left[M_{\psi(\cdot)}\big((\alpha-\frac{\sigma\gamma}{\tau})\mathbf{w}^\star - \frac{\beta\sigma\sqrt{\delta}}{\tau}\,\mathbf{h}, \frac{\sigma}{\tau}\big)\right],\ \text{as } p\to\infty. \tag{54}$$

We now appeal to Lemma 9 in Appendix A of (Thrampoulidis et al., 2018), which allows us to replace the Moreau envelope with their expected value due to the convergence we are getting in (54). Hence, by replacing the Expected value of the Moreau envelope function, we are getting the following optimization, to be analyzed in the next section.

$$\min_{\sigma\geq 0,\alpha} \max_{\substack{\gamma\\ \beta\geq 0,\tau>0}} \frac{1}{p}\,\mathbb{E}\left[M_{\psi(\cdot)}\big((\alpha-\frac{\sigma\gamma}{\tau})\mathbf{w}^\star - \frac{\beta\sigma\sqrt{\delta}}{\tau}\,\mathbf{h}, \frac{\sigma}{\tau}\big)\right] + \beta\sqrt{c_\kappa(\alpha,\sigma)} - \frac{\sigma\tau}{2} - \frac{\sigma}{2\tau}(\delta\beta^2+\gamma^2\kappa^2). \tag{55}$$

[3] The square-root trick: This is adopted from (Thrampoulidis et al., 2018), where it was proposed in the analysis of the auxiliary optimization in regularized M-estimators, and the idea is to use the following equivalence (which is derived immediately from AM-GM inequality):

$$\sqrt{x} = \min_{\tau>0}\ \frac{1}{2\tau}x + \frac{\tau}{2}\,,\ \ \forall x>0.$$

### A.3. Optimality condition of the auxiliary optimization

In this section, we conclude the proof of the main result of the paper by showing that (when $\delta > \delta^*$) the optimizer to the scalar optimization (55) can be derived by solving the nonlinear system of equations (11).
Here, we investigate the optimality condition for the solution of the auxiliary optimization. In Section A.2, we simplified the (AO) and after some algebra we got the scalar optimization (55) with respect to five variables. Here, we would like to present the solution to this optimization. Let $C(\alpha,\sigma,\gamma,\beta,\tau)$ denote the objective function in the scalar optimization. In other words, the fucntion $C$ is defined as:

$$C(\alpha,\sigma,\gamma,\beta,\tau) = \frac{1}{p}\,\mathbb{E}\left[M_{\psi(.)}\big((\alpha-\frac{\sigma\gamma}{\tau})\mathbf{w}^\star + \frac{\beta\sigma\sqrt{\delta}}{\tau}\,\mathbf{h}, \frac{\sigma}{\tau}\big)\right] + \beta\sqrt{c_\kappa\big(\alpha,\sigma\big)} - \frac{\sigma\tau}{2} - \frac{\sigma}{2\tau}(\delta\beta^2+\gamma^2\kappa^2). \tag{56}$$

The following lemma describes the behavior of the function $C$ with respect to its variables.

**Lemma 5.** *The function $C:\mathbb{R}^5\to\mathbb{R}$ defined in (56) is (jointly) convex with respect to the variables $(\alpha,\sigma)$, and (jointly) concave with respect to the variables $(\gamma,\beta,\tau)$.*

The proof of this lemma is provided in Appendix C. Using the result of Theorem 1, the objective function, $C$, will diverge when $\delta<\delta^*$. For $\delta>\delta^*$ Lemma 5 states that the function $C$ is convex-concave. The following remark indicates that the optimal solution of the optimization problem does not happen at the boundry values.

**Remark 6.** *We need to show that the optimal solution does not happen at the boundary, i.e., at $\beta=0$, or $\sigma=0$. Taking the derivative with respect to $\beta$ at the objective function in (50), we will have $\frac{\partial}{\partial\beta}|_{\beta=0} = \sqrt{c_\kappa(\alpha,\sigma)} > 0$. Therefore, the optimal $\beta$ is nonzero. It can also be seen in the same optimization program that when $\sigma=0$, $\beta$ can choose its value arbitrary large and the optimal value would be $+\infty$. Hence, the optimal $\sigma$ is also nonzero as we have a minimization w.r.t. $\sigma$.*

Let $(\bar\alpha,\bar\sigma,\bar\gamma,\bar\beta,\bar\tau)$ denote the solution to the optimization (55). Since the objective function is smooth with respect to its variables and the optimal values do not coincide with the boundries, its solution must satisfy the first-order optimality condition, i.e., $\nabla C\big(\bar\alpha,\bar\sigma,\bar\gamma,\bar\beta,\bar\tau\big) = \mathbf{0}_{5\times1}$. We will show that this would simplify to our system of nonlinear equations (11). We start by setting the derivative with respect to $\alpha$ to zero. We have,

$$\frac{\partial C}{\partial\alpha} = 0 \Rightarrow \frac{1}{p}\,\mathbb{E}\left[\frac{\partial}{\partial\alpha}M_{\psi(.)}\big((\alpha-\frac{\sigma\gamma}{\tau})\mathbf{w}^\star + \frac{\beta\sigma\sqrt{\delta}}{\tau}\,\mathbf{h}, \frac{\sigma}{\tau}\big)\right] + \frac{\beta}{2\sqrt{c_\kappa\big(\alpha,\sigma\big)}}\cdot\frac{\partial c_\kappa\big(\alpha,\sigma\big)}{\partial\alpha} = 0\,, \tag{57}$$

where we used the Leibniz integral rule to bring the derivative inside the expectation. Using the result of Lemma 7, we can write the following,

$$\frac{1}{p}\,\mathbb{E}\left[\frac{\partial}{\partial\alpha}M_{\psi(.)}\big((\alpha-\frac{\sigma\gamma}{\tau})\mathbf{w}^\star + \frac{\beta\sigma\sqrt{\delta}}{\tau}\,\mathbf{h}, \frac{\sigma}{\tau}\big)\right] = \frac{\kappa^2\alpha\tau}{\sigma} - \kappa^2\gamma - \frac{\tau}{p\sigma}\,\mathbb{E}\left[\mathbf{w}^{\star T}\mathrm{Prox}_{\frac{\sigma}{\tau}\psi(.)}\big((\alpha-\frac{\sigma\gamma}{\tau})\mathbf{w}^\star + \frac{\beta\sigma\sqrt{\delta}}{\tau}\,\mathbf{h}\big)\right]. \tag{58}$$

Replacing (58) in (57) gives the following nonlinear equation,

$$\frac{\tau}{p\sigma}\,\mathbb{E}\left[\mathbf{w}^{\star T}\mathrm{Prox}_{\frac{\sigma}{\tau}\psi(.)}\big((\alpha-\frac{\sigma\gamma}{\tau})\mathbf{w}^\star + \frac{\beta\sigma\sqrt{\delta}}{\tau}\,\mathbf{h}\big)\right] = \frac{\kappa^2\alpha\tau}{\sigma} - \kappa^2\gamma + \frac{\beta}{2\sqrt{c_\kappa\big(\alpha,\sigma\big)}}\cdot\frac{\partial c_\kappa\big(\alpha,\sigma\big)}{\partial\alpha} \tag{59}$$

Next, we find another optimality condition by setting the derivative with respect to $\beta$ to zero. We have,

$$\frac{\partial C}{\partial\beta} = 0 \Rightarrow \frac{1}{p}\,\mathbb{E}\left[\frac{\partial}{\partial\beta}M_{\psi(.)}\big((\alpha-\frac{\sigma\gamma}{\tau})\mathbf{w}^\star + \frac{\beta\sigma\sqrt{\delta}}{\tau}\,\mathbf{h}, \frac{\sigma}{\tau}\big)\right] + \sqrt{c_\kappa\big(\alpha,\sigma\big)} - \frac{\sigma\delta}{\tau}\beta = 0\,, \tag{60}$$

Similar to (58), we can compute the expected derivative of the Moreau envelope function by appealing to Lemma 7,

$$\frac{1}{p}\,\mathbb{E}\left[\frac{\partial}{\partial\beta}M_{\psi(.)}\big((\alpha-\frac{\sigma\gamma}{\tau})\mathbf{w}^\star + \frac{\beta\sigma\sqrt{\delta}}{\tau}\,\mathbf{h}, \frac{\sigma}{\tau}\big)\right] = \frac{\beta\sigma\delta}{\tau} - \frac{\sqrt{\delta}}{p}\,\mathbb{E}\left[\mathbf{h}^T\mathrm{Prox}_{\frac{\sigma}{\tau}\psi(.)}\big((\alpha-\frac{\sigma\gamma}{\tau})\mathbf{w}^\star + \frac{\beta\sigma\sqrt{\delta}}{\tau}\,\mathbf{h}\big)\right]. \tag{61}$$

Replacing (61) in (60) will give the following nonlinear equation:

$$\boxed{\frac{1}{p}\,\mathbb{E}\left[\mathbf{h}^T\mathrm{Prox}_{\frac{\sigma}{\tau}\psi(.)}\big((\alpha-\frac{\sigma\gamma}{\tau})\mathbf{w}^\star + \frac{\beta\sigma\sqrt{\delta}}{\tau}\,\mathbf{h}\big)\right] = \sqrt{\frac{c_\kappa\big(\alpha,\sigma\big)}{\delta}}\,.} \tag{E2}$$

Next, we compute the derivative with respect to $\gamma$ and set it to zero. We have,

$$\frac{\partial C}{\partial \gamma} = 0 \Rightarrow \frac{1}{p}\,\mathbb{E}\,\big[\frac{\partial}{\partial \gamma} M_{\psi(.)}\big((\alpha - \frac{\sigma\gamma}{\tau})\mathbf{w}^\star + \frac{\beta\sigma\sqrt{\delta}}{\tau}\,\mathbf{h}, \frac{\sigma}{\tau}\big)\big] - \frac{\sigma\kappa^2\gamma}{\tau} \;= 0\;, \tag{62}$$

$$\frac{1}{p}\,\mathbb{E}\,\big[\frac{\partial}{\partial \gamma} M_{\psi(.)}\big((\alpha - \frac{\sigma\gamma}{\tau})\mathbf{w}^\star + \frac{\beta\sigma\sqrt{\delta}}{\tau}\,\mathbf{h}, \frac{\sigma}{\tau}\big)\big] = \frac{\sigma\kappa^2\gamma}{\tau} - \kappa^2\alpha - \frac{1}{p}\,\mathbb{E}\,\big[\mathbf{w}^{\star T}\mathrm{Prox}_{\frac{\sigma}{\tau}\psi(.)}\big((\alpha - \frac{\sigma\gamma}{\tau})\mathbf{w}^\star + \frac{\beta\sigma\sqrt{\delta}}{\tau}\,\mathbf{h}\big)\big]. \tag{63}$$

Replacing (63) in (62) will give the following nonlinear equation:

$$\boxed{\frac{1}{p}\,\mathbb{E}\,\big[\mathbf{w}^{\star T}\mathrm{Prox}_{\frac{\sigma}{\tau}\psi(.)}\big((\alpha - \frac{\sigma\gamma}{\tau})\mathbf{w}^\star + \frac{\beta\sigma\sqrt{\delta}}{\tau}\,\mathbf{h}\big)\big] = \kappa^2\alpha\;.} \tag{E1}$$

Also, replacing (E1) in the nonlinear equation (59) gives the following nonlinear equation:

$$\boxed{\frac{\partial c_\kappa(\alpha,\sigma)}{\partial \alpha} = \frac{2\kappa^2\gamma}{\beta}\sqrt{c_\kappa(\alpha,\sigma)}\;.} \tag{E4}$$

Next, we take the derivative with respect to $\sigma$. We have:

$$\frac{\partial C}{\partial \sigma} = 0 \Rightarrow \frac{1}{p}\,\mathbb{E}\,\big[\frac{\partial}{\partial \sigma} M_{\psi(.)}\big((\alpha - \frac{\sigma\gamma}{\tau})\mathbf{w}^\star + \frac{\beta\sigma\sqrt{\delta}}{\tau}\,\mathbf{h}, \frac{\sigma}{\tau}\big)\big] + \frac{\beta}{2\sqrt{c_\kappa(\alpha,\sigma)}}\frac{\partial c_\kappa(\alpha,\sigma)}{\partial \sigma} - \frac{\tau}{2} - \frac{1}{2\tau}(\delta\beta^2 + \gamma^2\kappa^2) = 0, \tag{64}$$

We use the result of the Lemma 7 to compute the derivative of $M_\psi(\cdot,\cdot)$ with respect to $\sigma$. We have,

$$\frac{1}{p}\,\mathbb{E}\,\big[\frac{\partial}{\partial \sigma} M_{\psi(.)}\big((\alpha - \frac{\sigma\gamma}{\tau})\mathbf{w}^\star + \frac{\beta\sigma\sqrt{\delta}}{\tau}\,\mathbf{h}, \frac{\sigma}{\tau}\big)\big] = \frac{1}{2\tau}\big(\gamma^2\kappa^2 + \delta\beta^2 + \frac{\alpha^2\kappa^2\tau^2}{\sigma^2} - \frac{\tau^2}{p\sigma^2}\,\mathbb{E}\,\Big\|\mathrm{Prox}_{\frac{\sigma}{\tau}\psi(.)}\big((\alpha - \frac{\sigma\gamma}{\tau})\mathbf{w}^\star + \frac{\beta\sigma\sqrt{\delta}}{\tau}\,\mathbf{h}\big)\Big\|^2\big) \tag{65}$$

Replacing this into (64) will give the following equation,

$$\frac{1}{p}\,\mathbb{E}\,\Big\|\mathrm{Prox}_{\frac{\sigma}{\tau}\psi(.)}\big((\alpha - \frac{\sigma\gamma}{\tau})\mathbf{w}^\star + \frac{\beta\sigma\sqrt{\delta}}{\tau}\,\mathbf{h}\big)\Big\|^2 = \alpha^2\kappa^2 + \frac{\beta\sigma^2}{\tau\sqrt{c_\kappa(\alpha,\sigma)}}\frac{\partial c_\kappa(\alpha,\sigma)}{\partial \sigma} - \sigma^2 \tag{66}$$

Similarly, by taking the derivative with respect to $\tau$, we have:

$$\boxed{\frac{1}{p}\,\mathbb{E}\,\Big\|\mathrm{Prox}_{\frac{\sigma}{\tau}\psi(.)}\big((\alpha - \frac{\sigma\gamma}{\tau})\mathbf{w}^\star + \frac{\beta\sigma\sqrt{\delta}}{\tau}\,\mathbf{h}\big)\Big\|^2 = \alpha^2\kappa^2 + \sigma^2} \tag{E3}$$

We can now simplify (66) to get the following equation:

$$\boxed{\frac{\partial c_\kappa(\alpha,\sigma)}{\partial \sigma} = \frac{2\tau\sqrt{c_\kappa(\alpha,\sigma)}}{\beta}} \tag{E5}$$

Finally, we make a change of variable by replacing $\tau$ with $\frac{1}{\tau}$ in the equations (E1), (E2), (E3),(E4), and (E5) will respectively give the desired equations in the system of nonlinear equations (11) as the optimality condition on the solution of the optimization (55). This concludes the proof.

## B. Proof of Theorem 1

In this section we prove the result presented in Theorem 1 which identifies the phase transition on the separability of the data. To this end, we exploit the result of Lemma 4 which associates the following optimization to the GMM optimization (36).

$$\begin{aligned} \min_{\substack{\alpha\in\mathbb{R}\\ \tilde{\mathbf{w}}\in\mathbb{R}^p\\ \tilde{\mathbf{w}}\perp\mathbf{w}^\star}} \quad & \frac{1}{p}\psi(\alpha\mathbf{w}^\star+\tilde{\mathbf{w}}) \\ \text{s.t.} \quad & \frac{1}{p}(\mathbf{h}^T\tilde{\mathbf{w}})^2 \geq n\cdot c_\kappa\big(\alpha,\frac{\|\tilde{\mathbf{w}}\|}{\sqrt{p}}\big). \end{aligned} \tag{67}$$

We first show that, as $p,n\to\infty$, $\delta>\delta^*=\delta^*(\kappa)$ is the necessary and sufficient condition for the optimization program (67) to have a feasible solution. Define $\sigma:=\|\tilde{\mathbf{w}}\|\sqrt{p}$, and write the following:

$$\sup_{\substack{\alpha\in\mathbb{R}\\ \tilde{\mathbf{w}}\perp\mathbf{w}^\star}} \frac{1}{p}(\mathbf{h}^T\tilde{\mathbf{w}})^2 - n\cdot c_\kappa\big(\alpha,\frac{\|\tilde{\mathbf{w}}\|}{\sqrt{p}}\big) = \sup_{\sigma\geq 0,\alpha} \sigma^2\cdot\left\|\mathbf{P}^\perp\mathbf{h}\right\|^2 - n\cdot c_\kappa\big(\alpha,\sigma\big)\,. \tag{68}$$

Note that we used the fact that the $\mathbf{P}^\perp$ is the projection onto the hyperplane orthogonal to $\mathbf{w}^\star$. The supremum is achieved iff $\tilde{\mathbf{w}}$ chooses its direction to be the same as $\mathbf{P}^\perp\mathbf{h}$. The optimization program has a feasible point if and only if the optimal value in (68) be nonnegative. In other words, the necessary and sufficient condition on the separability of the data is:

$$\exists\, r\geq 0,\ s\,,\ \text{ s.t. } r^2\cdot\left\|\mathbf{P}^\perp\mathbf{h}\right\|^2 - n\cdot c_\kappa\big(s,r\big)\geq 0 \iff \frac{1}{n}\left\|\mathbf{P}^\perp\mathbf{h}\right\|^2 \geq \delta^* = \inf_{s,r\geq 0}\frac{c_\kappa(s,r)}{r^2}. \tag{69}$$

Next we note that $\mathbf{h}$ has i.i.d. $\mathcal{N}(0,1)$ entries, therefore, SLLN asserts that,

$$\frac{1}{n}\left\|\mathbf{P}^\perp\mathbf{h}\right\| \xrightarrow{a.s.} \frac{p-1}{n}. \tag{70}$$

Therefore, as $n,p\to\infty$ with $\delta:=\frac{p}{n}$, the optimization program (67) is feasible if and only if $\delta>\delta^*$.As Lemma 4 states that the solution to the GMM optimization (36) converges in probability to the solution of (67). Therefore, $\delta>\delta^*$ indicates the phase transition for the existence of the GMM classifier.

We would also want to refer the interested reader to (Cover, 1965) for an astute geometric/combinatorial perspective on the phase transition behavior in binary classification.

## C. Proof of Lemma 5

Consider the objective function in the optimization program 50, i.e.,

$$f^{(p)}\big(\alpha,\sigma,\mathbf{u};\gamma,\beta,\tau,\mathbf{v}\big) = \frac{1}{p}\psi(\mathbf{u}) + \beta\cdot\sqrt{c_\kappa\big(\alpha,\sigma\big)} + \frac{1}{p}\mathbf{v}^T\big(\mathbf{u}-\alpha\mathbf{w}^\star\big) - \frac{\sigma\tau}{2} - \frac{\sigma}{2\tau}\cdot\left\|\frac{\beta}{\sqrt{n}}\mathbf{h} - \frac{1}{\sqrt{p}}\mathbf{v} + \frac{\gamma}{\sqrt{p}}\mathbf{w}^\star\right\|^2. \tag{71}$$

First, we would like to show that $f^{(p)}$ is jointly convex with respect to $\alpha$, $\sigma$ and $\mathbf{u}$. From Lemma (10), we know that $\sqrt{c_\kappa\big(\alpha,\sigma\big)}$ is jointly convex with respect to $\alpha$ and $\sigma$. The function $\psi(\cdot)$ is also convex and the remaining terms are all linear with respect to these three variables. Hence, $f^{(p)}$ is convex with respect to $\mathbf{u}$, $\alpha$ and $\sigma$.
Next, we show that this function is jointly concave with respect to the remaining variables. We note that the function $\left\|\frac{\beta}{\sqrt{n}}\mathbf{h} - \frac{1}{\sqrt{p}}\mathbf{v} + \frac{\gamma}{\sqrt{p}}\mathbf{w}^\star\right\|^2$ is convex with respect to variables $\mathbf{v}$, $\gamma$, and $\beta$. The perspective of this function $\frac{1}{\tau}\left\|\frac{\beta}{\sqrt{n}}\mathbf{h} - \frac{1}{\sqrt{p}}\mathbf{v} + \frac{\gamma}{\sqrt{p}}\mathbf{w}^\star\right\|^2$ is (jointly) convex with respect to $(\gamma,\beta,\tau,\mathbf{v})$. Therefore, $f^{(p)}$ is jointly convex with respect to these variables as the remaining terms are affine with respect to $(\gamma,\beta,\tau,\mathbf{v})$. Next, we define the function $C^{(p)}$ by maximizing $f^{(p)}$ with respect to $\mathbf{v}$, i.e.,

$$C^{(p)}\big(\alpha,\sigma,\mathbf{u};\gamma,\beta,\tau\big) = \max_{\mathbf{v}\in\mathbb{R}^p} f^{(p)}\big(\alpha,\sigma,\mathbf{u};\gamma,\beta,\tau,\mathbf{v}\big) \tag{72}$$

This function is also jointly convex-concave, since it is a point-wise maximum of concave function with repect to $\mathbf{v}$. The result is the consequence of the fact that $C^{(p)}$ converges to $C$, i.e.,

$$C^{(p)}\big(\alpha,\sigma,\mathbf{u};\gamma,\beta,\tau\big) \xrightarrow{P} C(\alpha,\sigma,\gamma,\beta,\tau)\,,\ \text{ as } p\to\infty. \tag{73}$$

## D. GMM for Various Structures

In this section, we provide some technical details on how to characterize the performance of the classifiers introduced in Section 4. For each of the three classifiers, depending on the distribution of the underlying parameter ($\mathbf{w}^\star$) we simplify the nonlinear system (11) by explicitly evaluating the expected values.

### D.1. Max-margin classifier ($\ell_2$-GMM)

As mentioned earlier, when $\psi(\cdot) = \frac{1}{2}\|\cdot\|_2^2$, the GMM classifier will become the well-known max-margin classifier. In this case, we can find the following closed-form for the proximal operator:

$$\text{Prox}_{\frac{t}{2}\|\cdot\|^2}(\mathbf{v}) = \frac{1}{1+t}\mathbf{v}. \tag{74}$$

Therefore, the expectations in the nonlinear system (11) can be computed explicitly as follows:

$$\begin{cases} \frac{1}{p}\,\mathbb{E}\left[\mathbf{w}^{\star T}\text{Prox}_{\frac{\sigma\tau}{2}\|\cdot\|^2}\big((\alpha-\sigma\tau\gamma)\mathbf{w}^\star+\beta\sigma\tau\sqrt{\delta}\mathbf{h}\big)\right] = \dfrac{\kappa^2\big(\alpha-\sigma\tau\gamma\big)}{1+\sigma\tau}, \\ \frac{1}{p}\,\mathbb{E}\left[\mathbf{h}^{T}\text{Prox}_{\sigma\tau\psi(\cdot)}\big((\alpha-\sigma\tau\gamma)\mathbf{w}^\star+\beta\sigma\tau\sqrt{\delta}\mathbf{h}\big)\right] = \dfrac{\beta\sigma\tau\sqrt{\delta}}{1+\sigma\tau}, \\ \frac{1}{p}\,\mathbb{E}\left\|\text{Prox}_{\sigma\tau\psi(\cdot)}\big((\alpha-\sigma\tau\gamma)\mathbf{w}^\star+\beta\sigma\tau\sqrt{\delta}\mathbf{h}\big)\right\|^2 = \dfrac{\kappa^2\big(\alpha-\sigma\tau\gamma\big)^2+\beta^2\sigma^2\tau^2\delta}{\big(1+\sigma\tau\big)^2}. \end{cases} \tag{75}$$

Replacing these evaluations into the first three equations in the nonlinear system (11), will explicitly give two of the variables in terms of the other three variables. More specifically, we get $\gamma = -\alpha$ from the first equation, and $\beta = \frac{1+\sigma\tau}{\tau\sqrt{\delta}}$ from the thrid equation in the nonlinar system (11). Hence, the nonlinear system would reduce to solving the following system of 3 nonlinear equations with 3 unknowns:

$$\begin{cases} \sqrt{c_\kappa(\alpha,\sigma)} = \sigma\sqrt{\delta}, \\ \dfrac{\partial c_\kappa(\alpha,\sigma)}{\partial\alpha} = \dfrac{-2\kappa^2\alpha\tau\sigma\delta}{1+\sigma\tau}, \\ \dfrac{\partial c_\kappa(\alpha,\sigma)}{\partial\sigma} = \dfrac{2\sigma\delta}{1+\sigma\tau}. \end{cases} \tag{76}$$

### D.2. Sparse classifier ($\ell_1$-GMM)

The second choice for the potential function is $\psi(\cdot) = \|\cdot\|_1$, which is used to promote sparsity in the underlying parameter. Here, we assume that the entries of the underlying parameter are generated independently from the distribution $\Pi_s$ introduced in (27), where $s \in (0,1)$ denotes the sparsity factor which indicates the probability of an entry being nonzero. The nonzero entries have Gaussian distribution with variance $\kappa^2/s$. The proximal operator for $\ell_1$ norm can be computed explicitly as,

$$\text{Prox}_{t\|\cdot\|_1}(\mathbf{u}) = \eta(\mathbf{u},t), \tag{77}$$

where $\eta(x,t) = \frac{x}{|x|}\big(|x|-t\big)_+$ is the soft tresholding function that has been applied entrywise. The expectations that appear in the first three equations in the nonlinear system (11) can be presented as follows:

$$\begin{cases} \frac{1}{p}\,\mathbb{E}\left[\mathbf{w}^{\star T}\text{Prox}_{\frac{\sigma\tau}{2}\|\cdot\|^2}\big((\alpha-\sigma\tau\gamma)\mathbf{w}^\star+\beta\sigma\tau\sqrt{\delta}\mathbf{h}\big)\right] = \ 2\kappa^2\cdot Q(t_1)\cdot\big(\alpha-\sigma\tau\gamma\big), \\ \frac{1}{p}\,\mathbb{E}\left[\mathbf{h}^{T}\text{Prox}_{\sigma\tau\psi(\cdot)}\big((\alpha-\sigma\tau\gamma)\mathbf{w}^\star+\beta\sigma\tau\sqrt{\delta}\mathbf{h}\big)\right] = \big[2sQ(t_1)+2(1-s)Q(t_2)\big]\cdot\beta\sigma\tau\sqrt{\delta}, \\ \frac{1}{p}\,\mathbb{E}\left\|\text{Prox}_{\sigma\tau\psi(\cdot)}\big((\alpha-\sigma\tau\gamma)\mathbf{w}^\star+\beta\sigma\tau\sqrt{\delta}\mathbf{h}\big)\right\|^2 = 2\sigma^2\tau^2\big(\frac{s}{t_1^2}\cdot\chi(t_1)+\frac{1-s}{t_2^2}\cdot\chi(t_2)\big), \end{cases} \tag{78}$$

where $t_1$ and $t_2$ are defined as,

$$t_1 = \frac{\sigma\tau}{\sqrt{\frac{\kappa^2}{s}(\alpha-\sigma\tau\gamma)^2+\beta^2\sigma^2\tau^2\delta}}\,,\quad t_2 = \frac{1}{\beta\sqrt{\delta}}, \tag{79}$$

and the function $\chi : \mathbb{R} \to \mathbb{R}_+$ is defined as:

$$\chi(t) = \mathbb{E}[\big(Z - t\big)_+^2] = Q(t)\big(1 + t^2\big) - t\phi(t)\ , \tag{80}$$

where the random variable $Z$ in the above expectation have standard normal distribution, and $\phi(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2)$ denotes the density of the standard normal distribution. Replacing the computed expectations in (78) in the nonlinear system (11) gives the sparse nonlinear system presented in (30).
It is worth mentioning that the sparse nonlinear system (30) can be solved efficiently via iterative numerical methods. A main advantage of the sparse nonlinear system is that it has be presented in terms of the $Q(\cdot)$ function which can be computed quickly in most numerical softwares (e.g. MATLAB). For our numerical simulations in Section 5 we used an accelerated fixed-point iterative method to find the solution of the nonlinear system.

### D.3. Binary classifier ($\ell_\infty$-GMM)

The third and last choice of the potential function is the $\ell_\infty$ norm. In this case the potential function is defined as $\psi(\cdot) = p\,\|\cdot\|_\infty$[4]. The following lemma determines how to compute the proximal operator in this case.

**Lemma 6.** *Let $\mathbf{u} \in \mathbb{R}^p$ have i.i.d. entries from a distribution $\Pi$. Then, for $t > 0$, we have:*

$$\mathit{Prox}_{tp\|\cdot\|_\infty}\big(\mathbf{u}\big) = \mathbf{u} - \mathit{Prox}_{\lambda\|\cdot\|_1}(\mathbf{u}), \tag{81}$$

*where $\lambda$ is defined as,*

1. *for $t \le \mathbb{E}\,|W|$, $\lambda$ is the unique solution of $\mathbb{E}\left[\big(|W| - \lambda\big)_+\right] = t$.*
2. *for $t \ge \mathbb{E}\,|W|$, then $\lambda = 0$.*

In the following subsections, we use the result of Lemma 6 to compute the proximal operator for two different models (i.e., two different distributions on the entries of $\mathbf{w}^\star$.)

#### D.3.1. $\ell_\infty$-GMM WITH SPARSE PARAMETER

Here, we consider the case where the entries of $\mathbf{w}^\star$ are drawn independently from the distribution $\Pi_s$ defined in (27). Note that when we set $s$ to 1 this distribution will be the same as i.i.d. Gaussian entries. Hence, the result in this section can be applied to the non-sparse setting (when the underlying parameter has i.i.d. Gaussian entries.)
Using the result of Lemma 6, in this case the proximal operator can be computed as follows,

$$\mathrm{Prox}_{\sigma\tau p\|\cdot\|_\infty}\big((\alpha - \sigma\tau\gamma)\mathbf{w}^\star + \beta\sigma\tau\sqrt{\delta}\mathbf{h}\big) = (\alpha - \sigma\tau\gamma)\mathbf{w}^\star + \beta\sigma\tau\sqrt{\delta}\mathbf{h} - \mathrm{Prox}_{\lambda\sigma\tau\|\cdot\|_1}\big((\alpha - \sigma\tau\gamma)\mathbf{w}^\star + \beta\sigma\tau\sqrt{\delta}\mathbf{h}\big). \tag{82}$$

where $\lambda$ is defined in terms of the proxies $t_1$ and $t_2$ (defined in (79)):

1. If $\frac{s}{t_1} + \frac{1-s}{t_2} > \sqrt{\frac{\pi}{2}}$, then $\lambda$ is the unique solution of the following nonlinear equation:

$$2s \cdot \big[\frac{1}{t_1}\phi(\lambda t_1) - \lambda Q(\lambda t_1)\big] + 2(1-s)\big[\frac{1}{t_2}\phi(\lambda t_2) - \lambda Q(\lambda t_2)\big] = 1\ . \tag{83}$$

2. If $\frac{s}{t_1} + \frac{1-s}{t_2} \le \sqrt{\frac{\pi}{2}}$, then $\lambda = 0$.

Therefore, after finding the value of $\lambda$ by solving equation (83), the proximal operator which appears in the first three equations of the nonlinear system (11) can be written explicitly in terms of the proximal operator of the $\ell_1$ norm which was illustrated in Section D.2. Also, similar to the case of $\ell_1$-GMM, the expectations are written in terms of the functions $Q(\cdot)$, and $\phi(\cdot)$. Therefore, the solution to the nonlinear system can be found efficiently using numerical solvers.

[4]The multiplication by the dimension, $p$, is necessary to ensure that all the terms in the optimization have constant ($\mathcal{O}(1)$) order.

#### D.3.2. $\ell_\infty$-GMM WITH BINARY PARAMETER

Here, we consider the case where $\mathbf{w}^\star$ has i.i.d. entries with distribution $\Pi = \kappa \cdot \text{RAD}(\frac{1}{2})$. To simplify our presentation, we define the following proxy:

$$t_3 = \big(\frac{\alpha}{\sigma\tau} - \gamma\big) \cdot \kappa \,.$$

Using the result of Lemma 6, in this case the proximal operator can be computed as follows,

$$\text{Prox}_{\sigma\tau p\|\cdot\|_\infty}\big((\alpha - \sigma\tau\gamma)\mathbf{w}^\star + \beta\sigma\tau\sqrt{\delta}\mathbf{h}\big) = (\alpha - \sigma\tau\gamma)\mathbf{w}^\star + \beta\sigma\tau\sqrt{\delta}\mathbf{h} - \text{Prox}_{\lambda\sigma\tau\|\cdot\|_1}\big((\alpha - \sigma\tau\gamma)\mathbf{w}^\star + \beta\sigma\tau\sqrt{\delta}\mathbf{h}\big). \quad (84)$$

where $\lambda$ is defined as:

1. When $\beta\sqrt{\delta}\cdot\phi(-\frac{t_3}{\beta\sqrt{\delta}}) + t_3\cdot Q(-\frac{t_3}{\beta\sqrt{\delta}}) > \frac{1}{2}$, $\lambda$ is defined as the unique solution of the following equations:

$$\beta\sqrt{\delta}\cdot\phi(\frac{\lambda - t_3}{\beta\sqrt{\delta}}) + (t_3 - \lambda)\cdot Q(\frac{\lambda - t_3}{\beta\sqrt{\delta}}) = \frac{1}{2} \quad (85)$$

2. Otherwise, $\lambda = 0$.

Hence, $\lambda$ can be computed by solving the equation (85), and consequently the proximal operator which appears in the first three equations of the nonlinear system (11) can be written explicitly in terms of the proximal operator of the $\ell_1$ norm which was illustrated in Section D.2.

## E. Mathematical Tools

### E.1. Some useful lemmas

We gathered here some useful mathematical lemmas that are used in the proof of our main results. The following two lemmas are borrowed from (Salehi et al., 2019), and will be used to handle the Moreau envelope of the potential function. We refer the interested reader to (Jourani et al., 2014) for a detailed study of the properties of the Moreau envelope functions.

**Lemma 7.** *Consider the Moreau envelope of the function $\Phi : \mathbb{R}^d \to \mathbb{R}$, defined as:*

$$M_{\Phi(\cdot)}(\mathbf{v}, t) = \min_{\mathbf{x}\in\mathbb{R}^d} \Phi(\mathbf{x}) + \frac{1}{2t}\|\mathbf{x} - \mathbf{v}\|^2 \,. \quad (86)$$

*The derivatives of the function $M_{\Phi(\cdot)}(\cdot,\cdot)$ can be computed as follows:*

$$\frac{\partial M_{\Phi(\cdot)}}{\partial \mathbf{v}} = \frac{1}{t}\big(\mathbf{v} - \textit{Prox}_{t\Phi(\cdot)}(\mathbf{v})\big) \quad , \quad \frac{\partial M_{\Phi(\cdot)}}{\partial t} = -\frac{1}{2t^2}\big\|\mathbf{v} - \textit{Prox}_{t\Phi(\cdot)}(\mathbf{v})\big\|^2 \,, \quad (87)$$

*where $\textit{Prox}_{t\Phi(\cdot)}(\mathbf{v})$ is the unique solution of the optimization (86).*

**Lemma 8.** *Let $\Phi : \mathbb{R}^d \to \mathbb{R}$ be an invariantly separable funciton such that for all $\mathbf{x} \in \mathbb{R}^d$, $\Phi(\mathbf{x}) = \sum_{i=1}^d \phi(x_i)$ where $\phi(\cdot)$ is a real-valued function. Then, for all $(\mathbf{v}, t) \in \mathbb{R}^d \times \mathbb{R}_+$,*

$$M_{\Phi(\cdot)}(\mathbf{v}, t) = \sum_{i=1}^d M_{\phi(\cdot)}(v_i, t) \quad , \quad \textit{and} \quad \textit{Prox}_{t\Phi(\cdot)}(\mathbf{v}) = \begin{bmatrix} \textit{Prox}_{t\phi(\cdot)}(v_1) \\ \textit{Prox}_{t\phi(\cdot)}(v_2) \\ \vdots \\ \textit{Prox}_{t\phi(\cdot)}(v_d) \end{bmatrix}. \quad (88)$$

In the next lemma, we show that the Moreau envelope of a Lipschitz function is itself a Lipschitz function.

**Lemma 9.** *Let $\Phi : \mathbb{R}^d \to \mathbb{R}$ be an $L$-Lipschitz function. Then, $M_{\Phi(\cdot)}(\cdot, t)$ is a $2L$-Lipschitz function, i.e., for all $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$,*

$$|M_{\Phi(\cdot)}(\mathbf{u}, t) - M_{\Phi(\cdot)}(\mathbf{v}, t)| \le 2L\|\mathbf{u} - \mathbf{v}\| \,. \quad (89)$$

*Proof.* In order to show this result, we need to find an upper bound on the dericative of the Moreau envelope. For all $\mathbf{v} \in \mathbb{R}^d$ we have,

$$\begin{aligned} L \left\| \mathbf{v} - \text{Prox}_{t\Phi(\cdot)}(\mathbf{v}) \right\| &\geq \Phi\big(\mathbf{v}\big) - \Phi\big(\text{Prox}_{t\Phi(\cdot)}(\mathbf{v})\big) \\ &\geq \frac{1}{2t} \left\| \mathbf{v} - \text{Prox}_{t\Phi(\cdot)}(\mathbf{v}) \right\|^2 , \end{aligned} \tag{90}$$

where the first inequality is due to the $L$-Lipschitzness of the function $\Phi(\cdot)$, and the second inequality is derived from the fact that $\text{Prox}_{t\Phi(\cdot)}(\mathbf{v})$ is the solution to the optimization (86). This gives the following bound on the distance of the proximal operator to the underlying vector.

$$\left\| \mathbf{v} - \text{Prox}_{t\Phi(\cdot)}(\mathbf{v}) \right\| \leq 2tL. \tag{91}$$

We can now bound the derivative $\frac{\partial M_{\Phi(\cdot)}}{\mathbf{v}}$ as follows,

$$\left\| \frac{\partial M_{\Phi(\cdot)}}{\partial \mathbf{v}} \right\| = \frac{1}{t} \left\| \big(\mathbf{v} - \text{Prox}_{t\Phi(\cdot)}(\mathbf{v})\big) \right\| \leq 2L \ , \ \forall \mathbf{v} \in \mathbb{R}^d. \tag{92}$$

This concludes the proof. □

The following lemma provides some information on the summary functional $c_\kappa(\cdot,\cdot)$, which will be used later in Section A.3 to find the optimality condition for the solution of a scalar optimization.

**Lemma 10.** *The function $f(s,r) := \sqrt{c_\kappa(s,r)}$ is (jointly) convex in (s,r).*

*Proof.* First, note that for $\mathbf{x} \in \mathbb{R}^n$, the function $\mathbf{x} \mapsto \|(\mathbf{x})_+\|$ is a convex function as it can be written as a supremum of convex(linear) functions.

$$\|(\mathbf{x})_+\| = \sup_{\substack{\mathbf{u}\in\mathbb{R}^n_+ \\ \|\mathbf{u}\|\leq 1}} \mathbf{u}^T\mathbf{x}. \tag{93}$$

For $n \in \mathbb{N}$ define the function $f_\kappa^{(n)}(s,r)$ as:

$$f_\kappa^{(n)}(s,r) = \frac{1}{\sqrt{n}} \left\| \big(\mathbf{1}_n - s\kappa\mathbf{h}\mathbf{y} + r\mathbf{g}\mathbf{y}\big)_+ \right\|, \tag{94}$$

where $\frac{1}{\sqrt{n}}$ denote the all-one vector, $\mathbf{h}, \mathbf{g} \in \mathbb{R}^n$ have i.i.d. $\mathcal{N}(0,1)$ entries and $Y \sim \text{RAD}\big(\rho(\kappa\mathbf{h})\big)$. It is readily seen that $f_\kappa^{(n)}(s,r)$ is jointly convex with respect to $s$ and $r$ as it is a combination of a convex function and a linear function. Using the LLN, we also have that,

$$f_\kappa^{(n)}(s,r) \xrightarrow{P} f(s,r) = \sqrt{c_\kappa(s,r)}. \tag{95}$$

Therefore, $f(s,r)$ is a convex function as it is a point-wise limit of convex functions. □

### E.2. Convex Gaussian min-max theorem (CGMT)

Our analysis of the generalizaed margin maximizer optimization is based on the recently developed convex Gaussian min-max theorem (CGMT). As mentioned earlier in Section 1, the CGMT framework associates with a Primary Optimization (PO), a nearly-separable Auxiliary Optimization (AO), from which various properties of the primary optimization, such as the phase transition, can be investigated.

Let the (PO) and the (AO) problems be defined respectively as follows:

$$\Phi(\mathbf{G}) := \min_{\mathbf{w}\in\mathcal{S}_\mathbf{w}} \max_{\mathbf{u}\in\mathcal{S}_\mathbf{u}} \mathbf{u}^T\mathbf{G}\mathbf{w} + f(\mathbf{u},\mathbf{w}), \tag{PO}$$

$$\phi(\mathbf{g},\mathbf{h}) := \min_{\mathbf{w}\in\mathcal{S}_\mathbf{w}} \max_{\mathbf{u}\in\mathcal{S}_\mathbf{u}} ||\mathbf{w}||\mathbf{g}^T\mathbf{u} + ||\mathbf{u}||\mathbf{h}^T\mathbf{w} + f(\mathbf{u},\mathbf{w}), \tag{AO}$$

where $\mathbf{G} \in \mathbb{R}^{m\times n}, \mathbf{g} \in \mathbb{R}^m, \mathbf{h} \in \mathbb{R}^n$, $\mathcal{S}_\mathbf{w} \subset \mathbb{R}^n, \mathcal{S}_\mathbf{u} \subset \mathbb{R}^m$ and $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$. Denote by $\mathbf{w}_\Phi := \mathbf{w}_\Phi(\mathbf{G})$ and $\mathbf{w}_\phi := \mathbf{w}_\phi(\mathbf{g},\mathbf{h})$ any optimal minimizers in (PO) and (AO), respectively.

**Theorem 3** (CGMT)**.** *(Thrampoulidis, 2016) In (PO), let $\mathcal{S}_\mathbf{w}$, $\mathcal{S}_\mathbf{u}$, be convex and compact sets, and assume $f(\cdot,\cdot)$ is convex-concave on $\mathcal{S}_\mathbf{w} \times \mathcal{S}_\mathbf{u}$. Also assume that $\mathbf{G}$, $\mathbf{g}$, and $\mathbf{h}$ all have entries i.i.d. standard normal. The following statements are true (the probabilities are taken with respect to the randomness in $\mathbf{G}$, $\mathbf{g}$, and $\mathbf{h}$.),*

1. *for all* $\mu \in \mathbb{R}$*, and* $t > 0$*,*
$$\mathbb{P}(|\Phi(\mathbf{G}) - \mu| > t) \leq 2\mathbb{P}(|\phi(\mathbf{g}, \mathbf{h}) - \mu| \geq t) \, . \tag{97}$$

2. *Let* $\mathcal{S}$ *be an arbitrary open subset of* $\mathcal{S}_{\mathbf{w}}$ *and* $\mathcal{S}^c := \mathcal{S}_{\mathbf{w}}/\mathcal{S}$*. Denote* $\Phi_{\mathcal{S}^c}(\mathbf{G})$ *and* $\phi_{\mathcal{S}^c}(\mathbf{g}, \mathbf{h})$ *be the optimal costs of the optimizations in* (PO)*, and* (AO)*, respectively, when the minimization over* $\mathbf{w}$ *is now constrained over* $\mathbf{w} \in \mathcal{S}^c$*. If there exists constants* $\bar{\phi}$*,* $\bar{\phi}_{\mathcal{S}^c}$*, and* $\eta > 0$ *such that,*

   - $\bar{\phi}_{\mathcal{S}^c} \geq \bar{\phi} + 3\eta$ *,*
   - $\phi(\mathbf{g}, \mathbf{h}) < \bar{\phi} + \eta$*, with probability at least* $1 - p$ *,*
   - $\phi_{\mathcal{S}^c}(\mathbf{g}, \mathbf{h}) > \bar{\phi}_{\mathcal{S}^c} - \eta$*, with probability at least* $1 - p$ *,*

   *then,* $\mathbb{P}(\mathbf{w}_\Phi(\mathbf{G}) \in \mathcal{S}) \geq 1 - 4p$ *.*

In the asymptotic regime, we often appeal to the following corollary which is an immediate consequence of Part 2 in Theorem 3.

**Corollary 1** (Asymptotic CGMT)**.** *(Thrampoulidis, 2016) using the same notations and assumptions as in Theorem 3, suppose there exists constants* $\bar{\phi} < \bar{\phi}_{\mathcal{S}^c}$ *such that* $\phi(\mathbf{g}, \mathbf{h}) \xrightarrow{P} \bar{\phi}$*, and* $\phi_{\mathcal{S}^c}(\mathbf{g}, \mathbf{h}) \xrightarrow{P} \bar{\phi}_{\mathcal{S}^c}$*. Then,*
$$\lim_{n \to \infty} \mathbb{P}(\mathbf{w}_\Phi(\mathbf{G}) \in \mathcal{S}) = 1 \, . \tag{98}$$

For further reading on the subject, please refer to (Thrampoulidis et al., 2015; 2018).