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) ℓ_2-GMM that is the max-margin classifier, (2) ℓ_1-GMM which encourages sparsity, and (3) ℓ_∞-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 ℓ_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 δ^* 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 ℓ_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, δ, 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 ℓ_p norms, for example, are absolutely scalable.

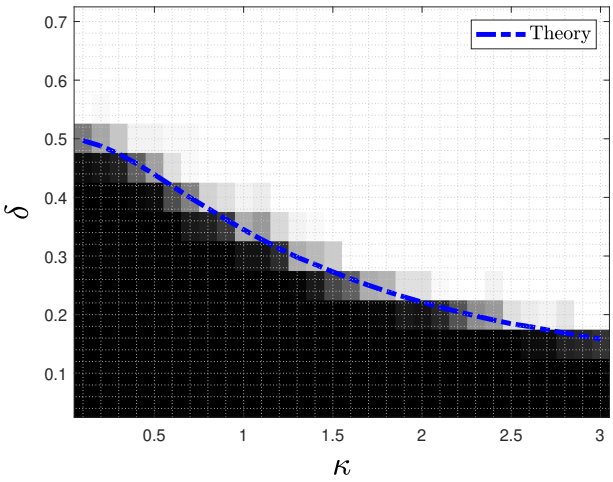

Figure 1: The phase transition, δ^*, for the separability of the dataset, where the feature vector, $\mathbf{x}_i$ is drawn from the Gaussian distribution, $\mathcal{N}(\mathbf{0}, \frac{1}{p}\mathbf{I}_p)$, and the labels are $y_i \sim \text{RAD}\big(\rho(\mathbf{x}_i^T\mathbf{w}^\star)\big)$, for $\rho(z) = \frac{e^t}{e^t+e^{-t}}$. 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 δ^* as κ varies.

Lemma 1. *δ^* is a decreasing function of κ, 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 Π 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 (ℓ_2-GMM)

The ℓ_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 ℓ_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 (β, and γ) 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 (ℓ_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 ℓ_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 ℓ_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 ϵ, 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 Π_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 (ℓ_∞-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 ℓ_∞ 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 ℓ_∞-norm can be derived by projecting the points onto the ℓ_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 Π, 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 λ is a non-increasing function of t, and $\lambda = 0$ whenever $t \geq \mathbb{E}\,|W|$.
Similar to the case of ℓ_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., ℓ_1, ℓ_2, and ℓ_∞ 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

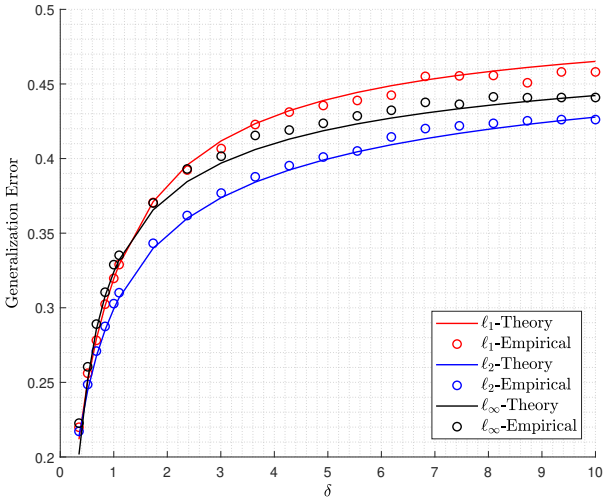

Figure 2: Generalization error of the general max margin classifier under three penalty functions, ℓ_1 norm with the red line (ℓ_1-GMM), ℓ_2 norm with the blue line (ℓ_2-GMM), and ℓ_∞ norm with the black line (ℓ_∞-GMM). **In this figure, the entries of $\mathbf{w}^\star$ are drawn independently from $\mathcal{N}(0,\kappa^2)$ 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 (ℓ_2-GMM) outperforms the other two classifiers. We should also note that as the overparameterization ratio, δ, grows the generalization error increases which indicates that the estimator is not reliable for large values of δ.
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

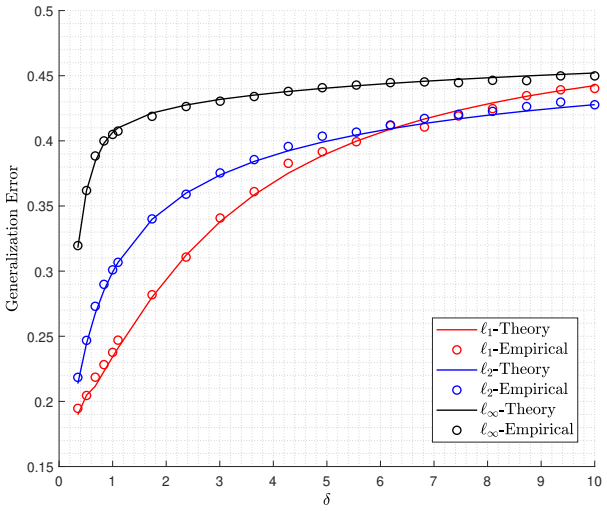

Figure 3: Generalization error of the general max margin classifier under three penalty functions, ℓ_1 norm with the red line (ℓ_1-GMM), ℓ_2 norm with the blue line (ℓ_2-GMM), and ℓ_∞ norm with the black line (ℓ_∞-GMM). **In this figure, the underlying vector $\mathbf{w}^\star$ is s-sparse, where the 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 $\kappa = 2$.

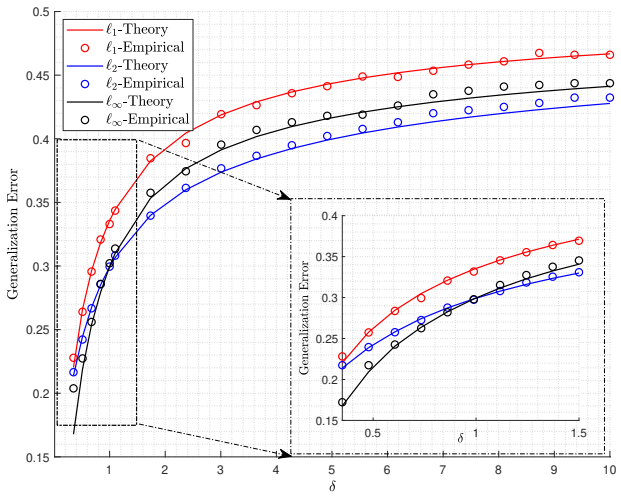

Figure 4: Generalization error of the general max margin classifier under three penalty functions, ℓ_1 norm with the red line (ℓ_1-GMM), ℓ_2 norm with the blue line (ℓ_2-GMM), and ℓ_∞ norm with the black line (ℓ_∞-GMM). **In this figure, the entries of $\mathbf{w}^\star$ are drawn independently from $\kappa * \textsc{Rad}(0.5)$ 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 ℓ_1-GMM outperforms the two other classifiers in the regime of δ that the classifiers performs well (i.e. $\delta \lessapprox 6$.) Similarly, we can observe that for large values of δ 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 ℓ_∞-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 δ.

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, ℓ_1, ℓ_2, and ℓ_∞ 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, ℓ_1-GMM outperforms the max-margin classifier (ℓ_2-GMM). We also observed that for binary signals, when $\delta < 1$, the ℓ_∞-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.